Université de Montréal

Prédiction des pré-miARN basée sur la conservation de structure dans les pri-miARN

par
Chabane Tibiche

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Juillet, 2005

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Prédiction des pré-miARN basée sur la conservation de structure dans les pri-miARN

présenté par:

Chabane Tibiche

a été évalué par un jury composé des personnes suivantes:

Sylvie Hamel
président-rapporteur

François Major
directeur de recherche

Julie Vachon
membre du jury

Mémoire accepté le 26 septembre 2005

RÉSUMÉ

Les microARN sont de petites séquences d'ARN d'environ 22 nucléotides. Leur principal rôle est la régulation génétique en bloquant la traduction de certains gènes en protéines. L'objectif de ce travail est d'identifier des séquences génomiques susceptibles d'être porteuses de microARN. Afin d'y arriver, nous nous sommes basé sur les séquences de microARN connus desquels nous avons extrait des caractéristiques structurelles et séquentielles que nous utiliserons pour établir les différents filtres qui composent notre méthode. Les filtres ont été choisis de manière à tenir compte de la séquence et de la structure. Les caractéristiques retenues sont celles relatives à la composition nucléotidique, l'énergie libre de repliement en structure secondaire, la différence d'énergie libre de repliement de la séquence native avec des séquences aléatoires de même composition nucléotidique globale et la conservation de la structure secondaire le long du processus de biosynthèse. Une attention particulière est accordée à la conservation de la structure secondaire. Cette dernière nous donne des résultats encourageants vu que nous avons considéré les séquences de manière à tenir compte de la composition nucléotidique aux voisinages immédiats des séquences porteuses de microARN ce que les travaux antérieurs ne considéraient pas. Au regard des résultats obtenus, cette caractéristique possède un important potentiel de discrimination et mérite une exploration plus approfondie.

Mots Clés : Micro-ARN, Régulation génétique, ARN non codant, interférence ARN, Régulation post-transcriptionnelle.

ABSTRACT

MicroRNAs are short endogenous sequences of about 22 nucleotides that are used to target messenger RNA for translational repression. This work is intended to predict genomic sequences of 65 ntds that may be putative microARN precursors. To do so, based on the known miRNA precursor sequences we derived some strucutural and sequencial features. These features are then used to build a prediction pipeline. The main carateristics considered are the secondary structure free folding energy, the differential free folding energy which is the difference between the free folding energy of the wild sequence and the mean of the free folding energies of the sequences resulting from the shuffling of the wild sequence, the nucleic composition and the conservation of the secondary structure of the precursor. A great attention is paid for the conservation of the secondary structure and the results show that it is worth to be considered since almost all of the known pre-miRNA have their structures well conserved. This has never been considered in previous miRNA prediction works eventhought it takes into account the nucleic composition in the neighborhood of the candidate pre-miRNA which influences the resulting secondary structure of the candidate precursor.

Keywords: Micro-RNA prediction, Gene regulation, post-transcriptional regulation, Gene repression, Gene Silencing, non coding RNA, RNA interference.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

NOTATION

A	: Adénine
ADN	: Acide DésoxyriboNucléique
ARN	: Acide RiboNucléique
C	: Cytosine
G	: Guanine
T	: Thymine
U	: Uracile
UTR	: UnTranslated Region
RISC	: RNA Induced Silencing Complex
RLC	: RISC Loading Complex

À ma femme Ouiza, pour son soutien indéfectible.
À mes deux enfants, Ianis et Rayan, qui ont insisté pour être partie prenante en me rappelant que la seule façon de décompresser c'est de jouer.
À mes amis, pour leurs précieux encouragements.

REMERCIEMENTS

Je tiens, avant tout, à remercier mon directeur, Dr. François Major, pour m'avoir accepté dans son laboratoire et pour son aide, ses orientations et conseils qui m'ont grandement aidés pour réaliser ce travail. Mes remerciements à tous les membres du LBIT qui m'ont facilité la tâche en répondant à mes nombreuses questions.

CHAPITRE 1

INTRODUCTION

Pendant longtemps, la biologie moléculaire faisait du caractère unidirectionnel de l'information génétique son dogme central soit l'ADN fait l'ARN et l'ARN fait la protéine. Le processus de synthèse va de l'ADN, support physique et stable de cette information, donc des gènes, vers les protéines, entités moléculaires assurant les réactions enzymatiques et autres fonctions régulatrices et/ou structurales indispensables à la vie de la cellule et donc des organismes vivants. Dans ce paradigme, les différents types d'acides ribonucléiques soient ARN messager (ARNm), l'ARN ribosomal (ARNr) et l'ARN de transfert (ARNt) sont perçus comme de simples intermédiaires impliqués dans le transport et le décodage de cette information génétique afin de produire des protéines.

Cette vision, centrée autour d'un rôle prépondérant des protéines, est en train de changer suite notamment à l'identification, il y une quinzaine d'années, d'un nombre croissant de gènes dits non codants [LFA93, CV01]. En effet, de tels gènes ne génèrent pas de l'ARN conventionnel mais plutôt de petits ARN, appelés microARN. Ces derniers, de tailles variables, assurent des fonctions cellulaires d'une extrême importance, notamment celles reliées au développement de certains organismes, sans pour autant être traduits en protéines [BHS+03, MLA97].

Le processus régissant la biosynthèse et le mode d'action de ces petits ARN, dont le recensement ne fait que débuter, s'avère être très minutieux et complexe [Bar04]. En ne regardant que les étapes inhérentes à ce processus, on peut être tenté de dire

que cela semble être simple. Toutefois, les aspects temporels et fonctionnels sont loin de l'être. Pour preuve, plus d'un millier de microARN ont été identifiés [EJ04] et de ce nombre seule une infime partie a sa fonction et son expression temporelle connues. Cette difficulté s'explique, entre autres, par l'incapacité des procédés biochimiques actuels à prendre en charge des expériences à l'échelle génomique à laquelle s'ajoute la nouveauté du phénomène de régulation génétique par des petits ARN endogènes [CV01].

Les premiers balbutiements de la régulation génétique par l'ARN endogène remontent aux débuts des années 90 quand Arsu et ses collègues ont démontré, pour la première fois en 1991 dans le ver *Caenorhabditis elegans*, l'implication du gène lin-4 dans la répression du gène lin-14 [AWR91]. Il a été constaté, quelques années auparavant, que le gène lin-14 est impliqué dans le développement du ver dans ses premiers stades larvaires [AH87].

En 1993, Ambros et ses collègues réussissent à isoler le gène lin-14 [LFA93] d'une longueur d'environ 22 nucléotides à son stade final de biosynthèse. La mutation ou la suppression des régions des gènes cibles, lin-14 et lin-28, complémentaires au miARN lin-4 mène à la perte de la fonction de régulation. Cette perte de la fonction de régulation se traduit par une stagnation du développement au même stade larvaire alors que la larve est sensée passer à un stade supérieur [OA99, WHR93, LFA93]. Ces études suggèrent que le gène lin-4 réprime les gènes lin-14 et lin-28 en se liant, avec une certaine imperfection, à des régions localisées dans les régions non traduites du côté 3' ou 3'UTR (*3' UnTranslated Regions*) de ces gènes. Les ARN messagers réprimés ne perdent pas leur stabilité [OA99] ce qui veut dire que la répression se fait par arrêt de la traduction et non par la destruction des ARNm

cibles comme il a été déjà observé dans l'action des petits ARN exogènes [FXM+98].

L'existence du petit ARN lin-4 dans le C. elegans était, au départ, perçue comme une caractéristique propre à ce ver. Cette perception a radicalement changé avec la découverte d'un autre gène de même nature, soit le gène let-7 [PRS+00]. Contrairement à lin-4, let-7 n'est pas propre au C. elegans mais se retrouve chez plusieurs autres espèces y compris l'humain, ce qui suggère que ce type de régulation est commun à un certain nombre d'organismes vivants. Cette suggestion a été confirmée par d'autres études qui ont révélé, plus tard, que ce phénomène est assez généralisé et qu'il est assuré par plusieurs petits ARN [LQRLT01, LA01, LLWB01] connus alors sous le nom de petits ARN temporels conséquemment à leurs tailles et modes d'expression mais qu'on appelle maintenant microARN ou miARN. Beaucoup de ces miARN sont conservés à travers le temps [OU05, TS04] et les études confirment que la régulation génétique par de petits ARN endogènes est commune à plusieurs espèces vivantes.

Les plantes ne sont pas en reste, d'autres miARN y seront découverts [LKRC02, MvdWMM02, Jon02, RWR+02]. Si les étapes de la biosynthèse des miARN chez les plantes sont semblables à ce qui a été observé chez les animaux, le mode d'action, lui, est différent. En effet, chez les plantes une complémentarité parfaite ou presque entre le miARN et sa région cible dans l'ARNm est nécessaire pour qu'il y ait catalyse [MW05, HWR96]. Cette catalyse se solde par la destruction de l'ARNm ce qui n'est pas le cas chez les animaux. Chez ces derniers, une complémentarité imparfaite particulière est suffisante pour arrêter la traduction alors qu'une complémentarité parfaite mène à la dégradation de la cible.

Les miARN ont été initialement associés au développement de certains organismes vivants [MLA97, WHR93, FA99, OA99] mais d'autres fonctions s'y ajouteront au fur et à mesure que les recherches avancent. Parmi ces fonctions, on retrouve la prolifération cellulaire [XGH04, XVGH03, BHS+03]. Ces nouvelles fonctions sont d'une importance capitale vu qu'elles peuvent être en relation avec certaines maladies tel que le cancer. Des signaux confirmant cette hypothèse ont commencé à apparaître il y trois années environ. Dans une première étude allant dans ce sens, Calin et ses collègues ont identifié un lien entre les miARN et la leucémie [CSD+04]. En effet, le gène codant pour les miARN *miR-15a* et *miR-16a* est inexistant dans les cellules atteintes de leucémie. Ces résultats suggèrent que la leucémie est peut-être due à l'absence de ce gène et donc des miRNA régulateurs qu'il contient. La ou les ARNm cibles de ces deux miARN ne sont pas encore identifiés mais il est fort probable que cette ou ces cibles soient impliquées dans la maladie. Dans une autre étude plus exhaustive, Calin et ses collègues sont arrivés à la conclusion que les miARN sont issus, de façon générale, des régions fragiles du génome notamment celles identifiées comme étant reliées au cancer [CSD+04]. Des résultats d'autres études vont de même sens [LGM+05, TKY+04, KTO+05]. Les miARN sont aussi impliqués dans le système de défense des cellules en ciblant certains rétrovirus [LDA+05].

Les miARN sont vulnérables. Dans une étude récente [LC04], Lu et Cullen ont montré que les miARN peuvent être la cible de certains adénovirus. Ils ont constaté que l'adénovirus VA1 agit sur l'exportation des pré-miARN vers le cytoplasme ainsi que sur l'activité de l'enzyme responsable de la maturation des miARN.

La compréhension plus approfondie de la régulation post-transcriptionnelle et de son niveau d'implication dans la vie des organismes vivants passent par l'identi-

fication des miARN, de leurs cibles ainsi que le degré d'implication de ces dernières dans la vie cellulaire. Aucune de ces trois tâches n'est aisée, toutefois, toute avancée, aussi petite soit elle, peut être d'une grande importance. C'est dans ce cadre que s'inscrit ce travail dans lequel nous essayons d'explorer d'autres avenues pour pouvoir identifier de nouveaux gènes de miARN.

La principale caractéristique bien intégrée dans les différents outils de prédiction de pré-miARN est la structure secondaire en forme d'épingle à cheveux commune à tous les pré-miARN déjà identifiés [EJ04]. Cette structure est un motif structural duquel on ne peut pas se passer. La principale raison est que c'est justement cette structure qui fait que le transcrit primaire est reconnu et catalysé [LAH+03, HLY+04, ZYC05]. Ce motif est aussi important pour l'agent d'exportation qui exporte le pré-miARN du noyau vers le cytoplasme [ZC04]. La seconde raison est l'absence d'autres signaux perceptibles propres aux pri-miARN ou aux pré-miARN ou encore aux miARN matures en mesure de nous aider à classer les candidats.

La structure de tige-boucle, même commune à tous les pré-miARN, n'est pas capable à elle seule de filtrer les bons candidats. Il est donc impératif de la jumeler à d'autres critères et c'est que nous avons fait dans ce travail. La méthode que nous proposons consiste, par une approche statistique, à attribuer un score aux séquences analysées. Ce score tient compte de la composition nucléotidique, de l'énergie libre de repliement, de la différence d'énergie libre de repliement de la séquence à analyser et du degré de conservation de la structure secondaire de cette dernière dans la structure secondaire d'une séquence de 500 nucléotides. Pour chacune de ses caractéristiques un score limite est calculé en se basant sur les pré-miARN connus. Le calcul de ses scores limites s'est fait à partir des précurseurs des

miARN connus. Les séquences qui n'auront pas surpassé un de ces scores limites seront rejetées. La nouveauté dans cette méthode est la prise en compte du transcrit primaire du miARN. Le transcrit primaire est simulé en prenant une séquence de 500 nucléotides ayant en son milieu le précurseur. Le repliement de cette longue séquence nous donne une multitude de structures sous-optimales et c'est à travers ces structures que la conservation de la structure du précurseur est recherchée. Cette methode nous donne une liste de séquences classées de la plus probable à la moins probable et nous estimons que les séquences ayant les meilleurs scores représentent sont des candidats miARN potentiels.

Pour faciliter la compréhension de ce travail, nous avons jugé bon d'expliquer ce qui gravite autour de la régulation génétique. Dans le premier chapitre de mémoire, nous nous sommes penchés sur la biologie cellulaire. Nous avons introduit les notions de base nécessaires à la compréhension de la biosynthèse des miARN. Dans le second chapitre, nous avons fait une revue, assez détaillée, de ce qui a été publié en relation avec la biosynthèse et le mode d'action des miARN. Ces connaissances sont nécessaires pour bien situer ce travail. Dans le 3ème chapitre, nous survolons les travaux qui ont été déjà réalisés dans le domaine de la prédiction des pré-miARN. Le 4ème chapitre est consacré à l'explication de la méthode et des différents paramètres que nous avons utilisés. Le dernier chapitre résume les calculs effectués, les résultats obtenus et leur analyse et nous finissons ce mémoire par une conclusion.

CHAPITRE 2

ÉTAT DE L'ART SUR LES ACIDES NUCLÉIQUES

2.1 Introduction

Les acides nucléiques ont été à l'origine isolés dans le noyau cellulaire et sont communs à tous les organismes vivants. Ils existent sous deux formes : les acides désoxyribonucléiques (ADN) et les acides ribonucléiques (ARN). L'ADN est une macromolécule biologique qui se trouve dans le noyau cellulaire. Elle est formée par des millions, voir des milliards, de nucléotides selon les espèces. Les nucléotides sont des monomères composés essentiellement d'un acide phosphorique, d'un sucre et d'une base azotée [Lew04] (voir figure 2.1).

Les nucléotides sont au nombre de cinq et forment deux classes selon la base azotée qu'ils contiennent. La classe des nucléotides à bases pyrimidiques composée de la Cytosine (C), la Thymine (T) et l'Uracile (U) et la classe des nucléotides à bases puriques composée de l'Adénine (A) et de la Guanine (G) [Eps03]. Contrairement aux autres nucléotides, l'Uracile ne se trouve que dans les ARN à la place de la Thymine qui, elle, ne se trouve que dans l'ADN. Les nucléotides sont des molécules polarisées et donc en mesure de former des liaisons hydrogènes lorsqu'ils sont mis en contact. L'Adénine et l'Uracile dans le cas de l'ARN, ou la Thymine dans le cas de l'ADN, se lient par deux liaisons hydrogènes. La Cytosine et la Guanine s'apparient par trois liaisons hydrogènes [Lew04]. Les nucléotides en mesure de s'apparier sont dits complémentaires.

Les nucléotides, par des liaisons chimiques, forment de très longues chaînes

nucléiques appelées brin d'ARN ou d'ADN . Ces liaisons, appelées ponts phosphodiesters, se produisent entre l'acide phosphorique du nucléotide $i+1$ et un des carbones du sucre du nucléotide i. À cause de la polarité de ses constituants, les brins polynucléotidiques d'ADN s'associent, selon les règles énoncées plus haut, pour former de l'ADN double brin. L'ADN double brin, tel qu'il est trouvé dans la cellule de la majorité des espèces, est formé de deux brins parfaitement complémentaires et antiparallèles appelés brin sens et brin antisens. Le sens de lecture de l'ADN est de l'extrémité 5' vers l'extrémité 3'. L'appellation 5' et 3' des extrémités d'une séquence d'acides nucléiques est due à une convention. Par cette convention, les cinq carbones du sucre sont indexés de 1' à 5' dans le sens des aiguilles d'une montre (voir figure 2.1). De cette numérotation est issue l'appellation des extrémités de la séquence d'ARN ou d'ADN. L'extrémité 5' se trouve du côté du carbone 5' du sucre et l'extrémité 3' du côté du carbone 3'. L'ADN double brin adopte une conformation spatiale hélicoïdale [WC53].

2.2 Les acides désoxyribonucléiques (ADN)

Hormis quelques virus pour lesquels l'ARN est le support génétique, l'ADN est le dépositaire de l'information génétique de presque toutes les espèces vivantes. Cette information génétique est nécessaire à tout organisme vivant du plus simple au plus complexe. Elle préserve la lignée des espèces comme elle constitue un moule pour la production de protéines et autres molécules dont les cellules vivantes ont besoin. On parle d'information génétique car les génomes, constitués à base d'ADN, ne jouent aucun rôle actif dans la formation des organismes vivants mais fournissent plutôt des copies des séquences nucléotidiques nécessaires à la synthèse des protéines et autres enzymes. Chaque cellule de l'organisme est en possession de tous les gènes [Lew04]. Cet aspect d'équité est assuré par la réplication de l'ADN. La

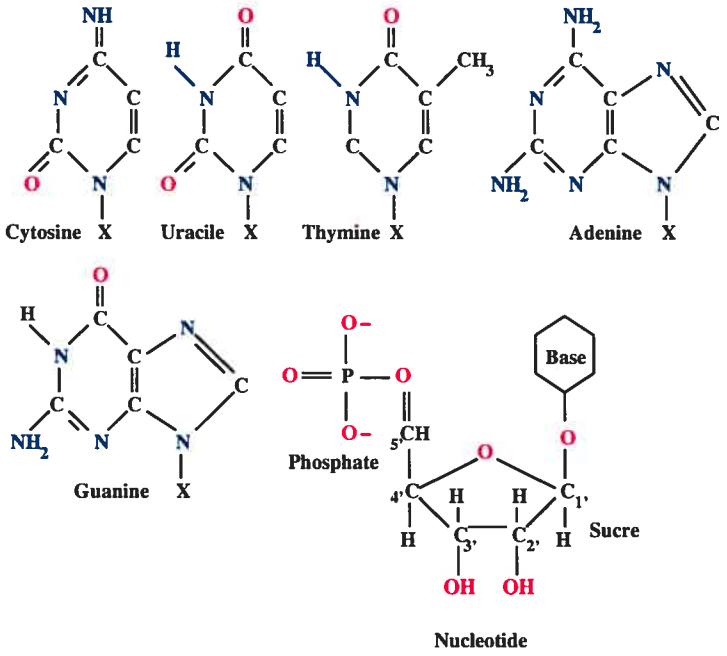

Figure 2.1 – Constituants des acides nucléiques.

réplication est un processus par lequel l'ADN est dupliqué dans la cellule parentale avant d'être transmise aux cellules filles lors la division cellulaire. La réplication de l'ADN n'est pas parfaite, des mutations se produisent parfois.

Physiquement, L'ADN peut être vu comme un ensemble de macromolécules, appelées chromosomes et, fonctionnellement, il peut être vu comme un ensemble d'unités fonctionnelles appelées gènes. Ces macromolécules se trouvent dans le noyau cellulaire. Chez l'humain, le génome est composé de 22 paires d'autosomes et de 2 chromosomes sexuels.

2.3 Les acides ribonucléiques (ARN)

L'ARN est une molécule polynucléotidiques mono-brin. Contrairement à l'ADN qui n'est qu'un support d'information, l'ARN est destiné à la formation d'unités actives

ou de complexes actifs formés par l'association de plusieurs de ces unités. Certains ARN, dits ARN codants, subissent des transformations avant d'acquérir la capacité d'être actifs tel que l'ARN messager qui est préalablement traduit en protéine. D'autres sont intégrés en tant que chaînes nucléotidiques dans des complexes tels que l'ARN ribosomal, l'ARN de transfert et d'autres petits ARN et sont appelés ARN non-codants.

2.3.1 Les ARN codants

Les ARN codants sont formés d'un seul type d'ARN soit l'ARN messager. Comme son nom l'indique, cet ARN a pour mission de transporter l'information nécessaire à la synthèse de protéines du noyau vers le cytoplasme ou la machinerie dédiée à la traduction y est domiciliée. Les transcrits primaires sont issus de l'ADN par un processus appelé transcription. Ces derniers sont épissés dans le noyau puis exportés vers le cytoplasme où ils seront traduits. Ci-dessous une description sommaire des étapes de biosynthèse des protéines.

2.3.1.1 Transcription

La transcription est un processus qui consiste à produire une copie d'une région de l'ADN. Seules certaines portions de l'ADN sont transcrites et elles sont appelées gènes. Les portions à transcrire présentent des signaux en amont, appelés promoteurs, et en aval qui signalent, respectivement, le début et la fin de la portion à transcrire [Rot93]. La duplication de la région à transcrire est réalisée par une enzyme appelée une polymérase d'ARN. La transcription passe par trois étapes appelées initiation, élongation et terminaison [Lew04].

1. Lors de l'initiation, le complexe enzymatique responsable de la transcription écarte les deux brins de l'ADN et via des protéines appelées facteurs de

transcription, reconnaît le promoteur et s'y fixe.

2. Lors de l'élongation, la polymérase d'ARN forme le brin d'ARN messager en appariant les nucléotides complémentaires au brin d'ADN à transcrire. Ces nucléotides se trouvent dans le nucléoplasme. Le double brin d'ADN se referme après le passage du complexe.
3. L'élongation s'arrête à la rencontre du signal de terminaison. L'ARN messager est libéré et les brins de l'ADN se referment.

Une modification est apportée aux transcrits originaux soit l'ajout d'une coiffe à l'extrémité 5' et d'une séquence polyadenine à l'extrémité 3' [Eps03]. Ces ajouts préviennent la dégradation du pré-ARNm par les exonuléases.

2.3.1.2 Epissage

Les transcrits primaires ne codent pas dans leur totalité pour des protéines. Ils contiennent des régions codantes, appelées exons, et des régions non-codantes, appelées introns, parsemées le long de la séquence [Lew04]. Ils contiennent aussi d'autres régions non-codantes appelées régions non-traduites aux extrémités 5' (5'UTR *UnTranslated Regions*) et 3' (3'UTR). Lors de l'épissage, les introns sont excisés et les exons recollés entre eux. Cette tâche est réalisée par des molécules appelées ribonucléoprotéines nucléaires. Les introns possèdent des signaux que le complexe d'épissage reconnaît. L'excision doit se faire à des endroits très précis et toute erreur fera perdre à l'ARNm final son sens. Une fois que l'épissage est réalisé, la séquence résultante est exportée du noyau vers le cytoplasme [Rot93].

2.3.1.3 Traduction

Cette étape consiste à traduire la séquence nucléique en une séquence polypeptidique. Cette opération est réalisée par une machinerie composée d'ARN ribo-

somaux, d'ARN de transfert et de quelques protéines. La traduction se fait par groupes de trois nucléotides appelés codons. En dehors des codons d'arrêt qui sont des signaux, chaque codon est traduit en un acide aminé selon un code universel. Comme il n'existe que 20 acides aminés naturels et qu'avec quatre nucléotides on peut former 4^3 soit 64 combinaisons de trois nucléotides, chaque acide aminé peut provenir de plus d'un codon. La traduction passe par trois étapes :

1. La première étape, appelée initiation, consiste à mettre en place le complexe traducteur. Cette mise en place fait intervenir certaines protéines appelées facteurs protéiques d'initiation. Ces facteurs reconnaissent le codon initiateur qui est presque toujours AUG.
2. La seconde étape consiste à lire les codons, à apporter les acides aminés correspondant et à les fixer à la chaîne polypeptidique en formation. L'apport des acides aminé est du ressort de l'ARN de transfert.
3. À la rencontre d'un des codons stop, la traduction s'arrête et la protéine est libérée.

Contrairement au codon initiateur qui code pour une Méthionine, les codons d'arrêt ne codent pour aucun acide aminé. Lors de la traduction, l'ARNm n'est pas détruit et peut donc servir à la synthèse d'une ou de plusieurs autres protéines.

2.3.2 Les ARN non-codants

Les ARN non-codants, comme le nom l'indique, ne codent pas pour des protéines mais sont plutôt intégrés en tant que chaînes nucléotidiques dans des complexes actifs. Les premiers à être caractérisés sont l'ARN ribosomal et l'ARN de transfert.

2.3.2.1 ARN ribosomal et ARN de transfert

L'ARN ribosomal ainsi que l'ARN de transfert sont, eux aussi, transcrits à partir de l'ADN dans le noyau cellulaire. Ils sont ensuite exportés vers le cytoplasme. L'ARN ribosomal, avec un certain type de protéines appelées protéines ribosomales forment le ribosome. Le ribosome représente l'unité de traduction des ARN messagers en protéines.

Les ARN de transfert, d'une centaine de nucléotides environ, sont eux aussi impliqués dans le processus de traduction. Leur rôle est de reconnaître le codon en instance de lecture et d'apporter l'acide aminé correspondant vers la chaîne polypeptidique. Les ARNt présentent une structure spatiale en forme de trèfle [Rot93]. Les ARNt possèdent deux sites d'une importance capitale, soient l'extrémité 3' et l'anticodon situé sur une des boucles. L'extrémité 3' présente trois nucléotides caractéristiques soient CCA-3'-OH [Rot93]. C'est sur ce site que se fixe l'acide aminé qui sera présenté au ribosome. L'anticodon doit s'apparier parfaitement avec le codon lu de l'ARN messager. L'appariement se fait par des liaisons hydrogènes et les triplets sont disposés de manière antiparallèle. Le code génétique de l'acide aminé que l'ARNt fixe à son extrémité 3' correspond au codon de l'ARN messager complémentaire à l'anticodon de cet ARNt.

2.3.2.2 Les petits ARN non-codants

On retrouve aussi d'autres petits ARN non-codants dans le noyau et le cytoplasme cellulaires. Certains de ces petits ARN ont été caractérisés, parmi lesquels on retrouve les microARN. Contrairement aux ARN précédents qui interviennent positivement dans le processus de synthèse des protéines, les microARN sont impliqués dans la répression de la traduction des ARNm qui est aussi connue comme la

régulation de l'expression génétique [LFA93]. D'autres outils sont connus pour être des régulateurs de l'expression génétique tels les promoteurs qui indiquent à l'ARN polymérase la région de l'ADN à transcrire [Lew04]. Les promoteurs interviennent dans le noyau pour empêcher la transcription d'une région de l'ADN alors que les microARN interviennent dans le cytoplasme pour arrêter la traduction.
La synthèse des microARN ressemble à celle des autres ARN. La transcription se fait à partir de l'ADN en des séquences de quelques centaines ou quelques milliers de nucléotides. Cette longue séquence nucléotidique et ensuite excisée pour ne laisser qu'une séquence d'environ 80 nucléotides de long structurée en épingle à cheveux. La prochaine étape est de l'exporter vers le cytoplasme où elle sera clivée une seconde fois pour donner naissance à un duplex d'environ 22 paires de bases. Ce duplex sera ensuite défait et un des deux brins sera recruté par un complexe actif de répression et il sera utilisé comme appât pour reconnaître les messagers à réprimer [Bar04].

2.3.3 Structures des acides ribonucléiques

Les nucléotides sont des molécules polarisées donc en mesures de former des liasons lorqu'ils sont mis en contact. L'Uracile et la l'Adenine se lient par deux ponts hydrogènes, la Guanine et la Cytosine se lient par trois ponts hydrogènes. Ces liaisons sont appelées liaisons Watson-Crick [Sae83]. La Guanine peut se lier à l'Uracile avec deux ponts hydrogènes et cette liaison est appelée liaison Wobble [Sae83].
Pour une séquence d'ARN plusieurs structures secondaires sont possibles et à chaque structure correspond une énergie libre. La structure la plus stable est celle qui minimise l'énergie libre. L'energie libre d'une structure et fonction des paires de bases et des boucles qui la composent 2.2. Les bases appariées favorisent la stabilité de la structure tandis que les boucles et les autres bases non appriées la

défavorisent. Les énergies libres des types de paires de bases (A :U, G :C, G :U) sont calculées en utilisant la thermodynamique et dependent principalement des liaisons hydrogènes dont est formée la paire de bases ainsi que des paires de bases voisines [KWT96, KBT99, XJB+98]. Certains algorithmes de prédiction de structures secondaires des ARN se basent sur la minimisation de l'énergie libre [Zuk03, PL88].

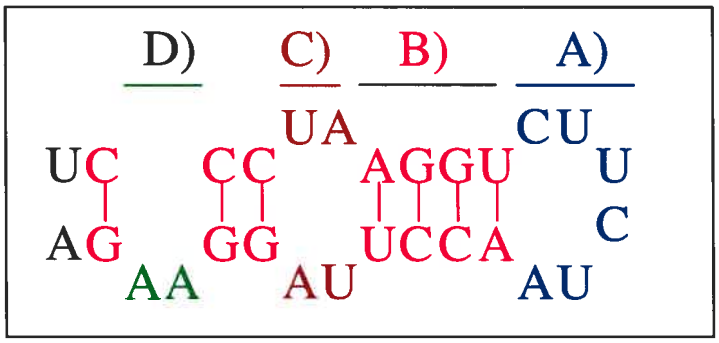

Figure 2.2 – Structure secondaire d'une séquence d'ARN. *L'énergie libre d'une structure secondaire est la somme des énergies d'appariement des paires de bases de l'helice* **(B)**. *La boucle terminale* **(A)** , *la boucle interne* **(C)** *et le bulge* **(D)** *défavorisent la stabilité de la structure secondaire*

CHAPITRE 3

MICROARN : BIOSYNTHÈSE ET MODALITÉS D'ACTION

3.1 Introduction

La régulation de l'expression des gènes est indispensable pour le bon fonctionnement cellulaire ainsi qu'au développement et au maintien de tout organisme vivant pluricellulaire. Des mécanismes génétiques, prévus à cet effet, permettent d'activer ou de réprimer l'expression d'un ou de plusieurs gènes sans toutefois altérer le support de l'information génétique. On peut distinguer quatre niveaux de régulation génétique :

- **Pré-Transcriptionnel :** Ce mode de régulation se fait par la désactivation du promoteur de transcription. Les promoteurs jouent le rôle d'indicateurs à l'ARN polymérase. En absence de ces promoteurs actifs sur l'ADN, cette enzyme n'est plus en mesure de reconnaître la région à transcrire [Lew04]. C'est ce qui est utilisé dans la spécialisation cellulaire car même si chaque cellule contient une copie complète du génome, et donc de tous les gènes, seule une partie de ces gènes a la possibilité de s'exprimer, ce qui donne une identité ou une spécialité à la cellule.
- **Post-Transcriptionnel :** Dans ce mode de régulation c'est la traduction qui est bloquée. Cette régulation est assurée par des miARN endogènes ou des petits ARN d'interférence (siARN) exogènes. Ces petits ARN, d'une longueur d'environ 22 nucléotides, s'intègrent dans des complexes ribonucléoprotéiques actifs qui les utilisent comme outils de reconnaissance des ARN messagers à réprimer. La régulation se manifeste soit par un simple arrêt de la traduction, sans effet sur la stabilité de la séquence du messager, ou par la destruction

de ce dernier.

- **Pré-Traductionnel :** Ce mode de régulation existe principalement chez les procaryotes. Chez ces organismes, en plus de codon de d'initiation de traduction soit le AUG, les ARN messagers sont dotés, en amont de ce dernier, d'une autre séquence AGGAGG appelée séquence de Shine-Dalgarno. L'absence de cette séquence affecte la traduction du messager [MAK02] indiquant sa probable implication dans la régulation de l'expression génétique.
- **Post-Traductionnel :** La régulation post-traductionnelle se fait sur les protéines. L'action est d'empêcher les protéines d'être activées en jouant sur la structure de ces dernières. C'est sur ce mode que se base la conception de la plupart des médicaments [Eps03].

La régulation génétique post-transcriptionnelle était encore inconnue il y a une quinzaine d'années [CV01]. C'est plus précisément en 1993 que Ambros et ses collègues ont pu isoler, dans le ver Caenorhabditis elegans, une séquence nucléotidique de 21 nucléotides, le microARN lin-4, responsable de la répression du gène lin-14 [LFA93]. Cette découverte a ouvert les portes à un grand champ d'exploration et plusieurs chercheurs s'intéressent de près à ce phénomène.

Les avancées, quoi que lentes, démontrent à quel point ce type de régulation est important car il est lié, entre autres, au développement des organismes. En effet, à titre d'illustration, le développement du ver C.elegans passe par trois stades larvaires avant d'atteindre le stade adulte. Le passage d'un stade à l'autre est assuré par un certain nombre de protéines dont fait partie lin-14 et lin-28, un autre gène régulé par lin-4 [MLA97]. L'expression de ces deux gènes se fait selon le stade dans lequel se trouve la larve. Au 1^{er} stade, les protéines nécessaires à

la vie du ver dans les autres stades ainsi que celles nécessaires au changement de stade larvaire sont inhibées, ne laissant que celles nécessaires au développement de l'organisme dans son premier stade. Au moment de passer du premier au second stade larvaire, il y a extinction des protéines utiles au premier stade pour laisser place à celles nécessaires au changement de stade. Cette activation et désactivation de protéines se fait par lin-4 à travers les gènes cibles lin-14 et lin-28 [MLA97]. La perturbation de ce processus de régulation de l'expression des gènes lin-14 et lin-28, par la mutation des sites d'intérêt du miARN lin-4 ou par l'inhibition de celui-ci, mène à la stagnation du développement du ver.

3.2 Biosynthèse des microARN

Succinctement, la biosynthèse des miARN est un processus minutieux qui peut être scindé en trois étapes. La première étape est la transcription. Elle consiste à donner naissance au transcrit initial du miARN appelé microARN primaire ou pri-miARN. Cette étape est semblable à la transcription de l'ARN messager. La seconde étape consiste à extraire, de cette séquence initiale, une sous séquence d'environ 80 nucléotides dont la seule spécificité avérée se trouve être la conservation de la structure secondaire en épingle à cheveux. Cette étape est analogue à l'épissage du pré-ARNm. Ces deux étapes, pour les deux types d'ARN, se déroulent dans le noyau de la cellule. La séquence de 80 nucléotides, appelée précurseur ou pré-miARN, est alors exportée vers le cytoplasme. Elle y sera clivée et le résultat sera un duplex d'ARN d'une longueur d'environ 22 paires de bases. De ce duplex, un des deux brins sera choisi et deviendra un miARN mature prêt à l'emploi.

3.2.1 Localisation

Les miARN peuvent être issus de toutes les régions du génome à des proportions différentes. La majorité des transcrits primaires sont situés dans les régions introniques des gènes codants pour des protéines (ARNc) ou dans de longs transcrits d'ARN non-codants (ARNnc).

Rodriguez *et al.* ont constaté que sur 232 miARN de mammifères analysés, 117 sont localisés dans les régions nommées plus haut [RGJAB04]. Approximativement 40% (90 miRNA) se trouvent dans des régions introniques des ARNc tels que mir-25, mir-93 et miR-106b qui sont localisés dans le gène hôte de MCM7, un gène codant pour une protéine impliquée dans la réplication de l'ADN [RGJAB04]. Environ 10% (27 miARN) se trouvent dans les régions introniques des ARNnc tel que miR-155 qui se trouve dans *BIC*, un transcrit précédemment identifié comme étant un ARNnc [Tam01]. On les retrouve toutefois dans d'autres régions du génome mais à de moindres proportions. Quelques uns se trouvent soient dans les introns ou dans les exons selon l'épissage alternatif du transcrit primaire du messager, d'autres sont localisés à cheval avec les régions exoniques des ARNnc [RGJAB04].

Dans leur travaux, Altuvia Y. *et al.* se sont intéressés, entre autres, aux distances entre les pré-miARN [ALL+05]. Ils ont pu constater que les miARN sont situés dans des agrégats. Sur les 207 miARN humains analysés, 37% sont situés dans des agrégats d'au moins deux éléments dont la distance est d'au plus 3 000 nucléotides. Sur ce point, nous considérons que prendre une longueur unique pour définir un agrégat pour tous les chromosomes est un peu subjectif. D'une part, parce que les chromosomes ne sont pas de même longueur, d'autre part, la densité de miARN varie d'un chromosome à l'autre.

Un autre facteur plus important dans la définition des agrégats est la grandeur des transcrits primaires générés pas l'ARN polymérase II. Car si ces agrégats existent, c'est, peut-être, pour que tous les miARN s'y trouvant soient transcrits au même temps et, dans ce cas, peut-on toujours parler d'agrégat si l'ARN polymérase II n'est pas en mesure de le transcrire en entier ? Une distance nucléique est plus en mesure de situer les miARN dans les chromosomes que de nous aider à définir des unités pouvant être transcrites.

3.2.2 Transcription

La transcription est la première étape dans la biosynthèse des miARN et elle a lieu dans le noyau cellulaire. Les transcrits primaires sont de longueurs variables allant jusqu'à quelques milliers de nucléotides et peuvent contenir un ou plusieurs miARN. Le processus de transcription des miARN est de mieux en mieux élucidé.

Ambros et ses collègues, dans leur travaux, ont constaté que le gène hôte du premier miARN connu, lin-4, est d'une longueur de 693 paires de bases [LFA93]. Lee Y. *et al.* [LKH+04] ont pu établir clairement, dans leur travaux, que les transcrits primaires des miARN sont des séquences de quelques centaines de nucléotides qui sont dotés d'une coiffe à l'extrémité 5' et d'une queue poly(A) à l'extrémité 3'. Ces caractéristiques indiquent que ces transcrits sont des produits de l'ARN polymérase II, une enzyme qui est aussi impliquée dans la transcription des ARN messagers. D'autres résultats d'expériences obtenus par la même équipe l'ont confirmé. En effet, le traitement de cellules humaines par α-amanitin, connu pour l'inhibition de l'ARN polymérase II, a conduit à la diminution de la concentration de pré-miARN dans les cellules traitées. Ceci met en évidence l'implication de l'ARN polymérase II

dans la transcription des miARN primaires. D'autres résultats vont dans le même sens [CHB04].

Kurihara *et al.* ont pu constater que chez les plantes, plus précisément chez l'*Arabidopsis Thaliana*, la reconnaissance du gène pri-miR-163 est rendue relativement facile par l'existence d'une boîte ou cassette TATA, semblable à ce que l'on trouve dans des ARN messagers [KW04]. Ces boîtes sont reconnues par l'ARN polymérase II lors de la transcription de ces derniers et que les deux pri-miR-163 obtenus par une amplification rapide de l'ADNc renferment un signal de polyadenylation. La présence de signaux semblables a été aussi observée dans le transcription des pri-miARN humain [CHB04]. Cette étude révèle l'existence d'un promoteur de transcription à l'extrémité amont et une polyadenylation à l'extrémité aval du transcrit primaire.

Ces évidences n'ont pas encore été explorées de manière plus approfondie sur les autres miARN connus mais il apparaît clairement que les transcrits primaires dédiés aux miARN sont de quelques centaines à quelques milliers de nucléotides et qu'il y aurait des signaux qui serviraient de guide à l'ARN polymérase II. Il n'est pas exclu que d'autres enzymes, telle que l'ARN polymérase III, soient impliquées notamment dans les pri-miARN originaires des régions introniques des ARN messagers [Bar04]. Car s'il est connu que l'ARN polymérase II est impliqué dans des transcriptions issues de l'ADN, des cas de transcription de miARN à partir d'un messager par l'ARN polymérase II n'ont pas encore été observés. Deux hypothèses émergent :

1. La première voudrait que le transcrit primaire soit destiné à devenir un ARN messager. Dans ce cas l'ARN polymérase II se charge de la transcription et lors de l'épissage certains des introns générés, de par leur structure, sont

prédestinés à devenir des miARN et seront traités comme tels en suivant le processus approprié de biosynthèse.

2. La seconde voudrait plutôt que le transcrit primaire soit destiné à donner naissance à des miARN. Dans ce cas aucun épissage n'a lieu et dans cette étape le flou reste entier. Car un transcrit primaire avec deux précurseurs ou plus est probablement soumis à une étape intermédiaire qui consisterait à découper la séquence initiale en petites séquences contenant chacune un précurseur, en quelque sorte une étape qui se substituerait à l'épissage. C'est dans ce cas que l'intervention d'autres ARN polymérases est envisageable.

3.2.3 Éboutage

Les transcrits primaires sont dans un premier temps éboutés pour n'en laisser qu'une séquence d'environ 70 nucléotides. La structure secondaire de la séquence résultante est une tige boucle et elle est conservée dans toutes les espèces [EJ04]. La tâche d'éboutage du pri-miARN incombe à *Drosha*, une enzyme de la famille des ARN polymérase III [LAH+03]. Cette famille peut être scindée en trois classes selon leurs compositions et l'organisation de leurs différentes composantes [HLY+04].

- La classe I, composée d'un domaine RNaseIII et d'une protéine liaison avec l'ARN double brin. On la retrouve dans les levures et les bactéries.
- La classe II, dont fait partie Drosha, est composée de deux domaines RNaseIII et d'un domaine de liaison avec l'ARN double brin. Cette classe ne se trouve que dans les animaux.
- La classe III renferme les homologues de Dicer, une enzyme impliquée dans la maturation des miARN.

Drosha, d'une grandeur de 130 à 160 kDa, présente à son N terminal une région riche en Proline précédée d'une autre région riche en Serine/Arginine, comme illustré dans la figure 3.1. Han et ses collègues ont pu mettre en évidence qu'en plus de la nécessité de la présence de toutes ses composantes, Drosha doit interagir avec une autre protéine, *DGCR8*, pour former un complexe fonctionnel capable de cliver la séquence primaire [HLY$^+$04].

Si le niveau d'activité du complexe n'est pas affecté par la suppression de toute la région riche en Proline jumelée à la suppression d'une partie de la région riche en Serine/Argenine, la suppression totale de ces régions particulières du N terminal altère l'association et le clivage. Les auteurs ont aussi confirmé que DGCR8 est indispensable à l'activité du complexe [HLY$^+$04]. Son rôle serait de stabiliser la liaison entre le double brin du transcrit primaire et le complexe comme il est possible que cette protéine joue un rôle dans l'orientation de la tige-boucle sur le complexe. Cette hypothèse est plus que probable puisqu'il est clairement établi que chacun des domaines RNaseIII clive un brin de la tige du pri-miRNA [HLY$^+$04]. Plus précisément RIIIDa s'occupe du brin 3' alors que RIIIDb agit sur le brin 5'. Donc il est évident qu'un ajustement doit être fait avant que la tige-boucle ne soit intégrée dans le complexe.

Zeng et ses collègues se sont intéressés à l'interaction de Drosha avec les pri-miRNA [ZYC05]. Ils ont confirmé que Drocha agit en un seul site et le résultat est une tige-boucle avec un débordement de 2 ou 3 nucléotides à l'extrémité 3' [LAH$^+$03, HLY$^+$04, ZKJ$^+$04a].

Il a été aussi constaté que le site de clivage de Drosha coïncide avec une des extrémités du miARN mature [LAH$^+$03, HLY$^+$04, ZYC05]. Si le miARN mature

est localisé sur le brin 5' de la tige, le clivage se fera à l'extrémité 5' du miARN mature et le brin 3' sera coupé à deux nucléotides en aval du vis-à-vis du site de clivage du brin 5. Si par contre la séquence mature est située sur le brin 3', le clivage du brin 3' se fait juste après le dernier nucléotides du miARN mature, le site de clivage du brin 5' est à deux nucléotides en aval. (voir la figure 3.2).

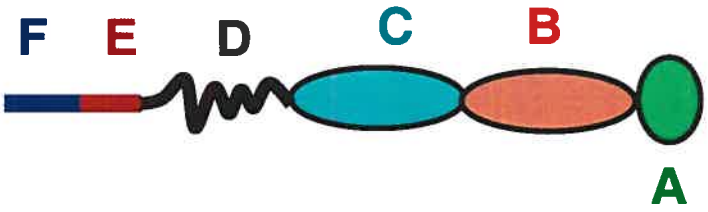

Figure 3.1 – L'ARN polymérase III Drosha. *Drosha est composé :* **(A)** *d'un domaine de liaison avec l'ARN double brin,* **(B)** *et* **(C)** *de deux domaines RNase III, RIIIDa et RIIIDb,* **(D)** *d'une séquence protéique intermédiaire,* **(E)**, *d'une région riche en Serine/Arginine et* **(F)** *d'une région riche en Proline*

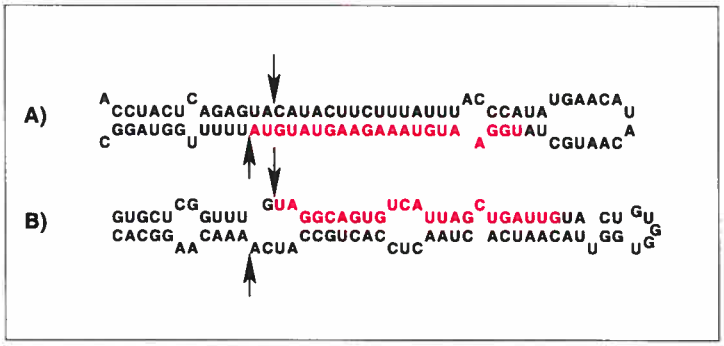

Figure 3.2 – Sites de clivage de Drosha. **(A)** *: Le miRNA mature est situé sur le brin 3'.* **(B)** *: Le miRNA mature est situé sur le brin 5'.*

Les opérations de clivage, quoi que assez caractérisées dans l'espace, ne le sont pas dans le temps donc on ne sait pas encore quel est le brin qui est traité en premier.

Cette étape du processus ne manque pas de soulever des questionnements sur ce que reconnaît Drosha dans la séquence mature. On sait que le site de clivage de Drosha coïncide avec une des extrémités du miARN mature mais on ne sait pas encore quel est le signal qui indiquerait ces extrémités. Car hormis d'être un peut plus riches en U comparativement aux autres régions, comme illustré dans la figure 3.3, les séquences matures ne recèlent pas d'autres caractéristiques notables. Considérant la minutie avec laquelle ces tâches doivent être exécutées, on ne peut pas dire que ces choix sont aléatoires. L'hypothèse qui nous vient est la possibilité de formation de liaisons hydrogènes entre les premiers nucléotides du miARN et un site du complexe et ces liaisons seraient favorisées par la présence de U comparativement aux autres nucléotides.

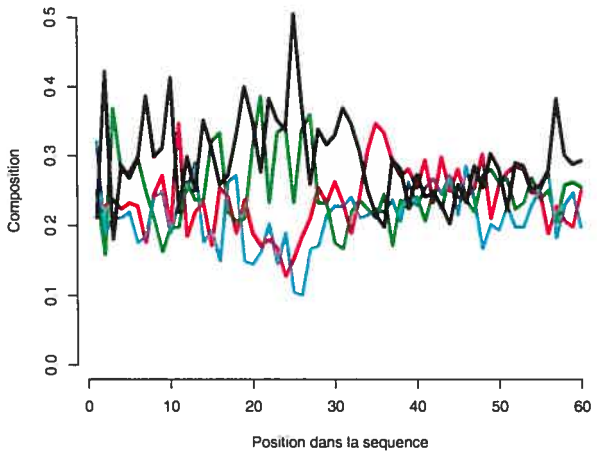

Figure 3.3 – Distribution des nucléotides le long des pré-miARN. *Les régions porteuses des miRNA matures, régions proches de l'extrémité 5' du pré-miARN, sont plus riches en U (courbe noire) comparativement aux autres nucléotides.*

Zeng et son équipe ont étudié l'effet de la longueur de la boucle terminale sur l'activité de Drosha. Il a été constaté qu'une boucle de moins de 9 nucléotides affecte considérablement l'activité de Drosha alors que des structures secondaires avec des boucles de 10 nucléotides ou plus sont bien transformées.

Comme Drosha traite des milliers, peut être des millions, de pri-miARN de compositions nucléotidiques différentes, il doit bien y avoir un signal que ce dernier reconnaît pour déterminer, avec une telle exactitude, son site de clivage. La présence d'une boucle commune à tous les miARN suggère une particularité fonctionnelle ayant un rôle important dans la catalyse qui a lieu à environ 22 nucléotides de cette dernière [ZYC05].

3.2.4 Exportation

Le transcrit primaire, une fois clivé dans le noyau par Drosha, est exporté vers le cytoplasme cellulaire. Cette tâche est assurée par Exportin-5 (Exp5) via le complexe du port nucléaire (NPC). Ces agents de transport sont des assemblages de macromolécules d'une grandeur variant de 50 kDa dans les levures à 160 kDa chez les mammifères [BCG04].

Exp5 se lie aux pré-miARN indépendamment de leurs séquences [BCG04] mais la structure semble jouer un rôle primordiale dans cette étape de la biosynthèse. En effet, si une absence de débordement à la base de la tige est tolérée, un débordement à l'extrémité 5' de la tige nuit considérablement à l'exportation du précurseur en inhibant la liaison de ce dernier à l'agent de transport [ZC04].

D'autres détails structuraux affectent l'exportation telles que la longueur de la boucle où les boucles courtes sont moins prisées par cet agent. De même, des struc-

tures secondaires avec des tiges courtes se lient moins bien à Exp5 et par conséquent s'exportent moins bien. Une fois dans le cytoplasme, ces structures restent stables impliquant que le processus de maturation n'est pas immédiat [BCG04]. En plus d'assurer le transport du noyau vers le cytoplasme, l'exportin-5 protège les miARN d'être altérés à leur arrivée dans le cytoplasme [ZC04].

3.2.5 Maturation

La maturation est la dernière étape dans le processus de biosynthèse des miARN. Elle consiste à cliver la tige-boucle à quelques paires de bases de la boucle terminale et d'en libérer un duplex d'ARN d'environ 22 paires de bases avec des débordements de 2 nucléotides aux deux extrémités 3' des deux brins du duplex. Cette tâche est dévolue à Dicer, une ARN polymérase III de la classe III.

Dicer est constitué de plusieurs domaines. En plus des deux domaines d'ARNase III *a* et *b* et du domaine de liaison avec l'ARN double brin, on y trouve un domaine PAZ (Piwi/Argonaute/Zwille), d'une protéine appelée DUF283, dont la fonction demeure inconnue, et d'un domaine ATPase/helicase [ZKJ+04b] (voir figure 3.4). PAZ est un complexe protéique propre à Dicer et à la famille Argonaute et est très conservé à travers les espèces. D'après des expériences réalisées avec des petits ARN d'interférence exogènes (siRNA), le domaine PAZ reconnaît le débordement de 2 nucléotides de l'extrémités 3' de la tige vue que ce domaine est connu pour son affinité pour de l'ARN simple brin.

Les deux nucléotides du débordement sont attirés vers une cuvette où ils sont emprisonnés. Les deux nucléotides suivants sont attirés vers un site de la protéine riche en résidus aromatiques et hydrophobes où des liens hydrogènes se forment

entre les acides aminés de ce site de PAZ et les oxygènes des liens phosphodiester du siARN [MYP04]. Il est fort probable que c'est ce domaine, PAZ, qui reconnaît les produits de Drosha, vu son affinité pour le débordement à l'extrémité 3' de ces derniers. En effet, le domaine de liaison reconnaît les ARN double brin via les débordements et s'y lie. Les ARNaseIII *a* et *b* s'accrochent chacun à un brin pour former un complexe stable capable de réaliser l'excision qui se fait à environ 22 nucléotides de la base de la tige.

S'il est connu, chez Drosha, que chacun des domaines *a* et *b* s'occupe d'un brin on ne sait pas encore s'il existe, chez Dicer, une quelconque affinité entre un brin particulier avec un domaine particulier. Si c'est le cas, il serait plus que probable que le duplex résultant de Dicer subisse une autre opération avant de se retrouver dans le complexe actif de répression. Cette opération intermédiaire serait dédiée à la séparation du duplex avant qu'un des deux brins ne soit recruté. Dans l'absence de cette étape intermédiaire, Dicer doit être en possession de deux domaines qui se chargeraient de vérifier la stabilité de chacune des extrémités du duplex pour pouvoir en sélectionner un brin.

Au niveau de cette étape, une divergence apparaît entre l'humain et la drosophile. Cette dernière possède deux complexes Dicer soient DCR1 et DCR2. L'un est dédié au clivage des siARN tandis que l'autre s'occupe des miARN. Chez l'humain, et les mammifères en général, un seul complexe, Dicer, prend en charge les deux types d'ARNnc. Ce dernier semble avoir une affinité pour les ARN double brin avec un débordement localisé à l'extrémité 3' et son efficacité dépend de la composition de ce débordement.

Vermeulan *et al.* ont constaté que, dans les siARN, la composition nucléotidique des débordements n'affecte pas la longueur des résidus de clivage mais elle agit grandement sur l'efficacité de Dicer [VBR+05]. Les ARN double brin sont efficacement traités s'ils ont des débordements se terminant avec un A ou ayant un C à l'avant dernière position. Un C à la position terminale ou un A à l'avant dernière position réduisent l'efficacité de Dicer [VBR+05].

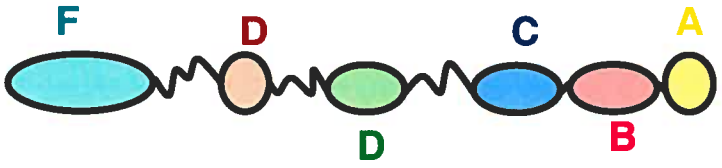

Figure 3.4 – L'ARN polymérase III Dicer. *Dicer est composé :* **(A)** *d'une protéine de liaison avec l'ARN double brin,* **(B)** *et* **(C)** *de deux domaines RNaseIII RIIIDa et RIIIDb,* **(D)** *d'un domaine PAZ,* **(E)** *de la protéine DUF283 et* **(F)** *et de ATPase/helicase*

3.3 Modalité d'action

Le résidu de Dicer, comme mentionné plus haut, est un duplex d'ARN avec des débordements aux deux extrémités 3' des brins. Ce duplex entrera dans une autre étape qui consiste à le défaire, de sélectionner un des deux brins et de l'intégrer dans un complexe ribonucléoprotéique actif qui a pour tâche de cibler les ARN messagers pour arrêter la traduction ou tout simplement pour les détruire. L'une ou l'autre des actions dépend du niveau de complémentarité du brin choisi et la région de l'ARN messager cible avec laquelle ce brin se lie.

3.3.1 Formation du complexe actif

Cette étape est bien modélisée chez la drosophile mais pas encore chez les mammifères. Elle consiste à prendre le duplex résultant de l'action de Dicer pour former

un complexe actif. En effet, l'activité des petits ARN d'interférence, siARN, passe par cette étape où le duplex est chargé dans un complexe intermédiaire appelé RLC (*RISC Loading Complexe*) pour finir dans un complexe de répression appelé RISC (*RNA-Induced Silencing Complexe*). Ce chargement se fait principalement via une interaction entre le duplex et R2D2, qui est une protéine avec deux domaines de liaison avec l'ARN double brin, qui interagit dans DCR2 lors du clivage. DCR2 est un des deux complexes Dicer dédié principalement au clivage des siARN double brin. L'autre complexe, DCR1, est dédié au clivage des pré-miARN [Tan05, LNP+04]. Rien n'indique que ce même procédé sera suivit par les miARN car ces derniers interagissent plutôt avec DCR1, qui ne possède qu'un seul domaine de liaison avec l'ARN double brin.

Le complexe de chargement des ARN d'interférence, s'initie par une interaction entre le domaine R2D2 est le duplex d'ARN. Les détails de cette interaction ne sont pas bien cernés, mais nous savons que R2D2 privilégie l'extrémité la plus stable [LNP+04], laissant la moins stable libre pour faciliter la séparation des brins du duplex. Cette asymétrie dans la sélection des brins caractérise aussi les complexes des miARN ou le brin choisi est celui dont l'extrémité 5' est la moins stable. Une fois dans le RLC, le duplex est défait et un des deux brins est sélectionné. Cette tâche, normalement dévolue pour une ARN helicase, semble faire appel à un ou plusieurs autres catalyseurs car cette opération ne se produit pas *in vitro*. Ce processus n'est donc pas aussi simple qu'il a été imaginé préalablement [Tan05].

Le complexe ainsi formé par le brin du siARN et R2D2, et probablement d'autres agents, est initialement inactif. Pour devenir actif, il s'associe avec une protéine de la famille des Argonautes qui a la capacité de cliver les ARN messa-

gers [Tan05, MLP+04]. Cette protéine, que l'on retrouve dans presque toutes les espèces, est souvent localisée dans le complexe RISC laissant sous-entendre un rôle dans le ciblage des ARNm.

Nous avons vu que c'est le domaine PAZ qui reconnaît les débordements de 2 nucléotides à l'extrémité 3' de la tige du précurseur et que ce domaine est composé de 3 protéines : Piwi, Argonaute et Zwille. Des résultats récents ont montré que l'extrémité 5' du brin choisi est emprisonnée dans une cuvette très conservée de la protéine Piwi [MYM+05]. Ce qui demeure méconnu c'est le comment et le quand de cette interaction. Piwi interagit dans une première étape avec le phosphate du bout 5'. En effet, pendant que le phosphate terminal est attiré dans la cuvette, les 4 ou 5 nucléotides suivants se lient à la protéine via les oxygènes du pont phosphodiester ayant la capacité de le faire. Cela concorde avec le calcul des surfaces enterrées de la protéine, où le brin actif couvre 1 500 $\AA^2$ tandis que le brin passif (ARNm) ne couvre que 350 $\AA^2$. Comme les séquences de la protéine Piwi de *Archaeoglobus fulgidus*, sur lequel ces tests ont été réalisés, et celles des Argonautes 1 et 2 humaines présentent des points de ressemblance, d'autres tests ont été réalisés en utilisant Argonaute2 humaine [MYM+05]. Une réduction de l'activité de clivage se fait sentir avec une mutation d'un acide aminé au niveau du site d'interaction entre Argonaute2 humaine et le brin d'ARN et elle est plus prononcée avec une mutation de deux acides aminés ce qui sous entend l'importance de ce site et de cette liaison.

L'association, chez l'humain entre Argonaute 2, la protéine d'excision, un brin du miARN, le brin directeur, et probablement d'autres protéines forme le complexe ribonucléoprotéique RISC prêt à agir est cibler les ARN messagers dans ses régions non-traduites.

3.3.2 Ciblage des ARN messagers

Le complexe actif ainsi formé cible les ARN messagers. L'action résultante dépend du degré de complémentarité entre le brin du miARN recruté par le complexe de répression et le messager ciblé. Une complémentarité parfaite conduit à la destruction de ce dernier alors que dans le cas d'une complémentarité imparfaite, le site actif ne pouvant probablement pas atteindre le brin du messager suite à une géométrie défavorable, ne fait qu'arrêter le processus de traduction de ce dernier [HZ02, MYM+05, SJA03].

Dans le cas d'une complémentarité imparfaite, le duplex formé par le petit brin et l'ARN messager doit satisfaire certaines conditions. La principale condition est l'appariement des 5 premiers nucléotides du côté 5' du miARN. Ce sont ces derniers qui sont présentés dans un premier lieu au messager. En présence d'un duplex instable, i.e un duplex qui n'a pas assez d'appariements entre ces premiers nucléotides éclaireurs, ce dernier se défait probablement tandis qu'en présence d'un duplex stable l'opération d'appariement a la possibilité d'aller de l'avant. Comme les 5 premiers nucléotides du miARN mature sont liés à Piwi [MYM+05], ils sont donc stables et adoptent probablement une conformation leur permettant de se lier, par des ponts hydrogènes, à l'ARN messager. C'est probablement la stabilité de ces premiers nucléotides qui favorise la formation d'autres appariements en aval. L'appariement du premier nucléotide n'est pas nécessaire et peut être même impossible à réaliser vu la conformation qui lui est imposée par l'emprisonnement du phosphate terminal dans la cuvette de la protéine Piwi.

Cette activité de clivage ou de répression sous-entend une accessibilité du miARN au messager [BCR05, OAF+05]. En absence de cette accessibilité, une diminution

drastique de l'activité du siARN a été constatée laissant croire que la complémentarité parfaite et globale, ou locale du côté 5', n'offre aucune certitude que le messager soit la cible du miARN ou siARN considéré. Cette complémentarité doit être prise dans un contexte plus globale en essayant d'intégrer le plus possible des éléments du messager cible, dont l'accessibilité. Dans le cas d'une complémentarité parfaite et d'un messager accessible ce dernier est dégradé par un clivage au niveau du 10^{eme} nucléotide [EMP^{+}].

Un certain flou règne encore quant à l'agent responsable de ce clivage. L'absence de tous les membres de la famille Argonaute mène à une absence totale de clivage, indiquant que c'est un membre de cette famille qui est responsable du clivage. D'autres possibilités commencent à émerger voulant qu'il y ait deux complexes RISC, un avec des capacités de clivage et l'autre sans. Le complexe doté d'une capacité de clivage dégrade le messager dans le cas d'une complémentarité parfaite et inhibe la traduction dans le cas d'une complémentarité imparfaite. Le complexe dépourvu de la faculté de clivage ne peut qu'arrêter la traduction même dans le cas d'une complémentarité parfaite [Tan05].

Cette possibilité de deux complexes soupçonnées chez la drosophile peut ne pas exister chez l'humain puisque on a constaté une divergence au niveau de la maturation. Cette divergence peut être le point de départ de deux voies, l'une pour les siARN avec complexe RISC pourvu d'une capacité de clivage et une autre pour les miARN sans cette dernière. Puisque chez les mammifères le clivage des siARN et miARN est réalisé par un seul complexe, Dicer, il est moins probable qu'il existe deux complexes RISC d'autant plus qu'aucune étape intermédiaire n'est encore mise en évidence.

Le site de clivage pour les siARN se situe entre les 10^{eme} et 11^{eme} nucléotides indépendamment de la longueur du siARN mature. Cela indique que le site de clivage dépend plus de la structure du complexe RISC dans l'espace que de la longueur ou la composition de siARN mature [EMP+].

3.4 Conclusion

Des différents résultats de recherche, nous avons pu constater l'importance de la structure secondaire des pré-miARN pour qu'ils soient recrutés par les différents complexes actifs. La structure seule semble expliquer la formation des complexes qui interviennent dans la biosynthèse des miARN.

Le premier motif structural est la boucle de l'épingle à cheveux. L'activité de Drosha dépend de la longueur de la boucle qui affecte aussi l'activité de l'agent d'exportation. Cela montre la concordance entre ces deux acteurs. Les débordements de 2 nucléotides à l'extrémité 3' de la structure sont aussi des motifs structuraux reconnus par l'agent de transport. La longueur de la tige affecte aussi l'efficacité de Exp5 où les tiges courtes sont moins bien exportées. Dicer, aussi, semble reconnaître les produits de Drosha par leurs débordements de 2 nucléotides à l'extrémité 3'. Pour le moment, cela semble être le seul motif notable en commun entre Drosha, Exp5 et Dicer, mais il n'est pas impossible que d'autres motifs structuraux locaux apparaîtront. Des motifs de séquence semblent aussi être utilisés par Drosha notamment pour la reconnaissance des extrémités du miARN mature lors du clivage du pri-miARN.

CHAPITRE 4

LE PROBLÈME ET LES TRAVAUX ANTÉRIEURS

4.1 Introduction

Les pré-miARN sont connus pour se replier en tige-boucle imparfaite [EJ04].Cette caractéristique, commune aux pré-miARN, est le point de départ dans la prédiction de ces derniers.

L'étape initiale d'identification de candidats potentiels par des outils bioinformatiques est nécessaire vu que des expériences biochimiques à la grandeur du génome avec les moyens actuels est impensable. En s'aidant de ce moyen, les chercheurs font des gains énormes en temps et en ressources. Toutefois les outils bioinformatiques ne peuvent pas, à eux seuls, nous prédire avec certitude ce genre de gènes ce qui rend indispensable une validation biochimique des candidats identifiés.

4.2 Définition du problème

Le problème est de trouver des séquences d'une centaine de nucléotides ayant des structures secondaires en forme d'épingle à cheveux susceptibles d'être des candidats potentiels pour devenir des gènes de miARN. Si trouver des séquences avec ce genre de structures secondaires est relativement facile et peu coûteux en temps, les filtrer et classer les candidats ne l'est pas. Nous voyons deux raisons à cela :

1. La première réside dans la divergence entre les compositions nucléotidiques de ces gènes. En effet, les alignements multiples sont inefficaces à produire une séquence consensus indiquant que la composition nucléotidique des précurseurs

n'est peut-être pas un critère important.

2. La seconde réside dans la divergence entre les structures. Les structures des pré-miARN ne présentent pas de motifs structuraux locaux notables en dehors de la structure globale en forme de tige-boucle.

Il est vrai que les outils actuels de prédiction de structures secondaires ne sont pas très fiables. Suite aux expériences qu'ils ont réalisés [KSW+04], Krol *et al.* ont constaté ces différences. Sur un ensemble de dix séquences, 8 montrent un certain nombre de différences de structure avec les prédictions de *mfold* [Zuk03]. Ces différences se trouvent surtout dans les boucles terminales et dans le nombre, la localisation et la longueur des bulges et des boucles internes [KSW+04]. Cela peut être d'une importance capitale sachant que la boucle terminale figure parmi les motifs structuraux très importants dans la biosynthèse des miARN [ZYC05, ZC04].

Il est possible qu'il y ait d'autres sites d'intérêt pour les enzymes et autres complexes intervenant dans la biosynthèse des miARN. C'est le cas pour la boucle terminale. Néanmoins d'autres motifs, même communs à un certain nombre de précurseurs, ne sont d'aucune utilité pour les complexes actifs et par conséquent inutiles pour la prédiction. L'identification des motifs structuraux importants ne peut pas se faire avec certitude au stade actuel de la recherche. Cela pousse la communauté scientifique à combiner les caractéristiques extraites des séquences de pré-miARN connus. La pondération de ces différentes caractéristiques se base principalement sur la prépondérance de ces dernières plutôt que sur leur apport dans le processus de biosynthèse. Cet apport ne peut être vérifié que par des expériences impliquant des délais et des ressources supplémentaires.

4.3 Survol des travaux antérieurs

Les outils de prédiction des miARN se basent, principalement, sur deux critères. Le premier est la structure secondaire en épingle à cheveux et le second est la conservation de la séquence entre espèces génétiquement proches. Chez les plantes, la condition d'un appariement presque parfait entre le miARN mature et ses cibles est aussi utilisée.

4.3.1 Chez les métazoaires

Peu d'outils bioinformatiques ont été conçus pour la prédiction des miARN pour les métazoaires. Le premier à être développé est MirScan. Il a été conçu initialement pour la prédiction des pré-miARN dans le ver C. elegans [LLW+03]. Cet outil se base sur la structure secondaire en forme d'épingle à cheveux, sur l'homologie de séquences avec *Caenorhabditis briggsae* et sur un ensemble d'autres motifs structuraux et séquentiels locaux.

Dans un premier temps les séquences conservées entre les deux espèces, et dont la structure secondaire est conforme aux critères, sont extraites. Cette première étape a pu retenir 50 des 53 pré-miARN déjà connus. Ces séquences serviront d'ensemble d'apprentissage pour établir une unité de mesure pour classer les autres candidats. Cette unité de mesure se base sur la conservation de la séquence et de la structure. Cette conservation tient compte de la conservation du résidu et de son état, apparié ou non. La séquence mature du miARN est aussi mise à contribution. Un niveau de conservation élevé est imposé à la région 5' du miARN mature comparativement à sa région 3'. La symétrie des boucles internes fait partie des critères ainsi que la distance entre la boucle terminale et le dernier nucléotide du miARN mature proche de celle-ci. Cette longueur varie de 2 à 9 nucléotides avec une prédominance des boucles de longueur de 4 à 6 nucléotides.

Des différentes caractéristiques considérées, la conservation de séquence et les appariements ont été les plus habiles à différencier les séquences d'apprentissages, donc connues, des autres séquences nouvellement extraites par homologie entre les deux espèces. Des 36 000 séquences extraites par homologie de séquences, 35 ont pu se classer parmi les plus probables en ayant des scores supérieurs à 13.9 qui est le score médian des précurseurs de l'ensemble d'apprentissage. De ces 35 candidats, 16 ont été confirmés par des expériences alors que 19 ne l'ont pas été et sont probablement des faux positifs. Au regard des résultats validés, il apparaît qu'il y a encore des critères à interférer pour la recherche des pré-miRNA.

Le second outil consiste à établir un ensemble de profils pour représenter les pré-miARN [LLG05]. Ces profils sont extraits après inspection visuelle des alignements de tous les pré-miARN connus. Ces profils ont servi d'entrées pour le logiciel ERPIN (http ://tagc.univ-mrs.fr/erpin/) et WU-BLAST (http ://blast.wustl.edu). ERPIN est un outil qui se base sur un alignement de séquences et d'une structure secondaire pour, dans un premier temps, créer un profil de structures secondaires qui sera utilisé comme requête à chercher dans une base de données de séquences. WU-BLAST, quant à lui, est un outil de recherche de séquences dans une base de données. Ces outils se basent principalement sur l'homologie de séquences. Ce fait les rend complètement inefficaces lorsqu'il s'agit d'identifier des pré-miARN propres à une seule espèce. Aucune expérience biochimique n'a été réalisée pour valider les 283 candidats rapportés comme potentiels par cette méthode.

Dans un même ordre d'idées, Weber [Web05], en se basant uniquement sur l'homologie de séquences entre l'homme, la souris et le rat, a identifié un certain nombre de candidats. La prédiction se base sur les précurseurs déjà existants dans

une espèce qu'il recherche dans les deux autres. Les séquences ayant un bon degré de similarité ainsi qu'une structure en forme d'épingle à cheveux sont rapportées comme des candidats pré-miARN.

Comme tous ces outils comptent parmi leurs critères la conservation de séquence et de structure entre espèces, ils sont incapables de nous fournir des candidats à partir d'une espèce, ce qui représente un sérieux handicap.

4.3.2 Chez les plantes

Contrairement aux outils dédiés à la prédiction des pré-miARN chez les métazoaires, findMiRNA est un outil qui ne fait pas appel à l'homologie de séquences [AJM+05]. Il fait intervenir, par contre, la complémentarité parfaite entre le miARN mature et sa cible. En effet, chez les plantes une complémentarité parfaite ou presque est requise pour qu'il y ait répression de l'ARN messager [LXKC02, PAW+03].
La méthode, dans son étape initiale, identifie les séquences génomiques en mesure de former des duplex avec une complémentarité presque parfaite avec des régions non-traduites de l'extrémité 3' de l'ARN messager de l'*Arabidopsis thaliana*. La méthode commence par identifier, dans les régions intergéniques, des séquences d'ancrages de longueur 7 qui ne doivent pas être des séquences répétées et au moins deux des nucléotides doivent être autres que A ou U. Ces séquences sont alors alignées avec les 3'UTR des messagers. Dans le cas d'un alignement sans erreurs de la séquence d'ancrage, une extension de l'alignement est faite. Chaque paire de base se voit attribuer un score qui est de 2 points pour un appariement Watson-Crick (AU et GC) et un point pour un appariement Wobble (GU). Les alignements parfaits de 18 à 25 paires de bases dont le score normalisé est d'au moins 1.55 et le score total d'au moins 35 points sont considérés et la séquence issue des régions intergéniques est considérée comme le miARN mature potentiel.

Des séquences de 400 nucléotides, ayant au milieu la séquence mature qui vient d'être identifiée, sont analysées pour d'éventuelles structures en épingle à cheveux. L'analyse se fait par glissement de la séquence mature en amont et en aval de son emplacement original et par attribution d'un score pour chaque alignement. Les scores unitaires sont de 2 points pour les appariements Watson-Crick, d'un point pour les appariements Wobble et de -1 pour toute paire de bases non appariée. L'alignement présentant le meilleur score est retenu pour la séquence de 400 nucléotides considérée. Le pré-miARN est la séquence délimitée d'un côté par une extrémité du miARN mature potentiel, identifié dans l'étape précédente, et de l'autre par une extrémité de la séquence ayant le meilleur score d'alignement avec le miARN mature potentiel. Ce candidat pré-miARN est ensuite replié en utilisant *RNAfold* [Hof03]. Une limite supérieure, en énergie, est imposée pour les candidats pour être retenus. Elle est calculée par la formule $E = -0.35 * l + 9.7$, où l est la longueur de la séquence.

D'autres méthodes intègrent la conservation de la séquence entres les espèces et d'autres caractéristiques. Wang et ses collègues [WRCG31], en plus de tenir compte de la conservation phylogénétique des séquences entre Arabidopsis thaliana et le riz, ont tenu compte de la composition nucléotidique, principalement le contenu en G+C, et les caractéristiques de la structure, notamment les longueurs de la tige et de la boucle et le nombre et la distribution des erreurs d'appariement le long de la tige.

Cette méthode commence, dans un premier temps, par identifier les séquences se repliant en tige-boucle. Cette étape consiste à prendre des séquences de 21 nucléotides et de rechercher des compléments en amont de ces dernières avec un taux d'erreurs d'alignement permis de 25% . Les deux séquences alignées peuvent se situer à une distance allant de 10 à 150 nucléotides comptés du dernier nucléotide de la séquence

requête au premier nucléotide de la séquence cible. En cas d'alignement valide, le précurseur sera la séquence allant du nucléotide situé à 20 nucléotides en amont de l'extrémité 5' de la séquence requête au nucléotide situé à 20 nucléotides en aval de l'extrémité 3' de la séquence cible. Chaque requête de 21 nucléotides et sa cible, séquence avec laquelle elle s'aligne, sont considérées comme des miARN potentiels. Ces séquences passeront ensuite par le filtre de la composition nucléotidique. Les séquences ayant des contenus en G+C plus petit que 30% ou plus grand que 70% seront rejetées. Les pré-miARN dont la longueur de la boucle varie de 20 à 75 nucléotides et ayant au moins une séquence mature seront retenues. Les miARN matures de 21 nucléotides sont ensuite alignées avec les régions intergéniques du génome du riz. Les séquences présentant des orthologues avec au moins 90% d'identité seront retenues. La dernière étape consiste à vérifier la structure des orthologues où l'absence de structures en tige-boucle entraîne le rejet du candidat.

CHAPITRE 5

PRÉSENTATION DE LA MÉTHODE DE PRÉDICTION DES PRÉ-MI-ARN

5.1 Introduction

Au regard de ce qui a déjà été fait dans la caractérisation et la prédiction des pré-miARN, un fait marquant en ressort. Dans la majorité des cas, les chercheurs se basent principalement sur la structure du précurseur et la conservation de la séquence de ce dernier entre les espèces génétiquement proches.

En ne considérant que le précurseur lors de la prédiction de la structure secondaire, on sous-entend que cette structure est indépendante de la composition nucléotidique dans ses environs immédiats. Ce qui n'est pas nécessairement le cas. La structure d'une séquence d'une centaine de nucléotides et la structure de cette même séquence prise dans une autre séquence de quelques centaines de nucléotides ne seront pas nécessairement identiques. Cela est dû principalement au fait que dans le premier repliement, chaque nucléotide a, au plus, quelques dizaines de partenaires avec lesquels il peut s'apparier alors que dans le second repliement, il a le choix entre quelques centaines de nucléotides, ce qui est non négligeable. Avec un tel choix, si la structure de la séquence du précurseur est conservée dans la structure de la grande séquence c'est parce qu'elle est probablement la structure la plus stable et donc possiblement un bon candidat pour être un précurseur de microARN.

La seconde caractéristique sur laquelle se base la prédiction des pré-miARN, soit la conservation de séquences entre espèces, sous-entend que la régulation post-

transcriptionnelle se fait par les mêmes gènes régulateurs, alors qu'il est possible que la divergence génétique ait induit une divergence dans ces gènes régulateurs, telles que les mutations. Il est vrai que ce qu'on observe en analysant les séquences connues supporte l'idée de la conservation de séquences mais il ne faut pas oublier que la majeure partie des pré-miARN connus ont été prédits en se basant sur la conservation de séquences entre les espèces proches. Cela n'est pas pour remettre en cause la validité du concept mais plutôt une tentative de tenir compte de la possibilité d'une divergence imposée par le temps.

5.2 Bases de données

Pour réaliser ce travail, nous avons eu à utiliser la base de données des micro-ARN (*miRNA registry*) [EJ04]. Cette base de données contient, dans sa version 6.0, 227 séquences de pré-miARN humains. Parmi les pré-miARN humains contenus dans cette base de données, on retrouve des séquences validées par des expériences biochimiques et d'autres qui ont été déterminées par des outils bioinformatiques ou par homologie de séquences avec d'autres espèces proches de l'humain, principalement le rat et la souris. Ces séquences ne correspondent pas tout à fait à ce qui se trouve dans la littérature, principalement les produits de Drosha. En effet, nous avons déduis de nos lectures que le clivage de Drosha coïncide avec une des extrémités de la séquence mature du miARN et que la structure secondaire présente un débordement de 2 ou 3 nucléotides à l'extrémité 3' par rapport à l'extrémité 5' [LAH+03, HLY+04, ZYC05], c'est à dire que le brin 3' de la tige est de 2 à 3 nucléotides plus long que le brin 5', ce qu'on n'observe dans aucune des structures présentes dans cette base de données.

Pour être conforme aux écrits scientifiques publiés, nous avons apporté des modifications à ces séquences tel qu'illustré à la figure 5.2. La modification des séquences

contenues dans la base de donnée de *miRNA registry* est faite dans le but de tenir compte uniquement des séquences résultantes de l'éboutage de Drosha (voir figure 5.1.

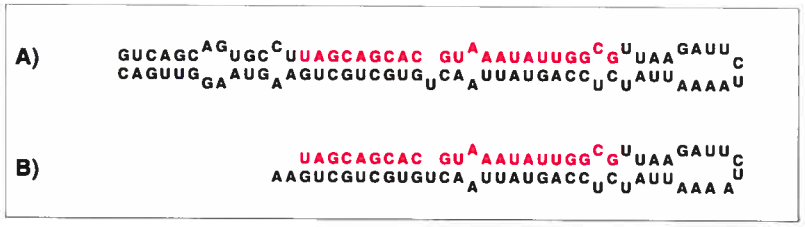

Figure 5.1 – Séquences de pré-miARN. *A) Séquence du pré-miARN hsa-mir-16-1 présent dans la base de données de miRNA registry. B) Séquence du même pré-miARN tel qu'ébouté par Drosha.*

L'homogénéisation des pré-miARN est faite comme suit :

- Si le miARN mature est situé à l'extrémité 5' du précurseur, nous supprimons la sous-séquence située à une distance de 2 nucléotides en amont du l'extrémité 5' de la séquence mature et la sous-séquence située à une distance de 3 nucléotides en aval du nucléotide vis-à-vis du premier nucléotide du miARN mature dans la structure du précurseur (voir figure 5.2-(b)).
- Si le miARN mature est situé à l'extrémité 3' du précurseur, nous supprimons la sous-séquence située à une distance de 2 nucléotides en aval de l'extrémité 3' du miARN mature et la sous-séquence située en amont du nucléotide vis-à-vis de l'avant-dernier nucléotide de la séquence mature. (voir figure 5.2-(a)).
- Si le précurseur porte 2 séquences matures, on applique une des règles précédentes si l'intégrité du miARN mature n'est pas touchée. Sinon la coupure se fait à un nucléotide en amont du miARN mature du brin 5' et un nucléotide en aval du miARN mature du brin 3'. (voir figure 5.2-(c)).

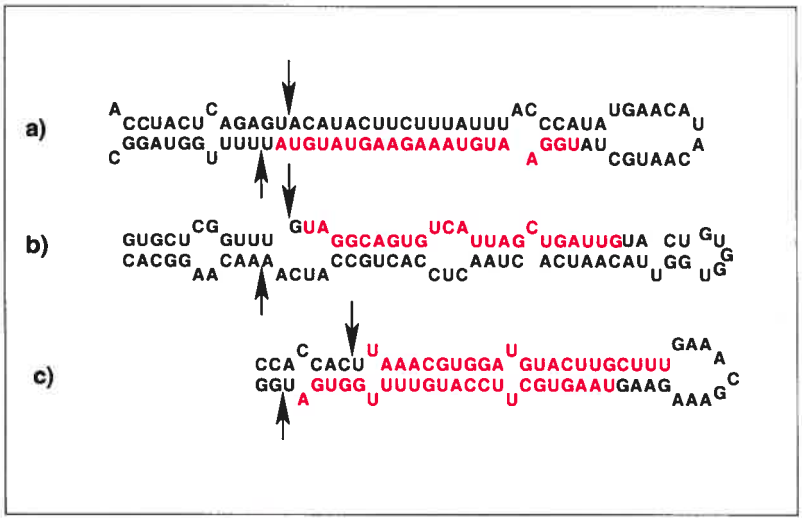

Figure 5.2 – Homogénéisation des pré-miARN

Sur les 22 pré-miARN contenant deux miARN matures, seuls deux se retrouvent dans le 3ème cas. Un des précurseurs présente une séquence mature du brin 3' qui dépasse de 5 nucléotides la séquence mature du brin 5' (*mir-302a*) et un autre avec un dépassement de 4 nucléotides (*hsa-mir-302c*). Les 20 précurseurs restants ont des débordements de 3 nucléotides ou moins, ce qui est en accord avec ce qui a été rapporté dans la littérature.

5.3 Méthode et application

Pour tenir compte des remarques formulées plus haut, nous avons développé une nouvelle méthode pour la prédiction de précurseurs de miARN qui se base sur l'application d'un certain nombre de filtres. Chaque filtre est relatif à une des caractéristiques considérées soient la composition nucléotidique, l'énergie libre de repliement, la différence d'énergie libre de repliement et la conservation de la structure secondaire. Les filtres prennent en entrée une séquence de 65 nucléotides et

nous donnent en sortie la séquence avec son score si celui depasse le score limite et rien sinon. Les filtres sont conçus pour tenir compte des deux principales caractéristiques des pré-miARN, soit la séquence et la structure.
Concernant la séquence, nous nous sommes basés sur la composition globale en A, C, G et U. Pour classer les structures, nous nous sommes basés sur l'énergie libre de repliement et sur la différence d'énergie libre de repliement entre la séquence native et la moyenne des énergies d'un certain nombre de séquences obtenues par un mélange des nucléotides de cette séquence native. L'énergie libre de repliement nous renseigne sur la stabilité de la structure tandis que la différence d'énergie libre nous donne une indication sur le degré de stabilité de la séquence native étant donnée la composition nucléotidique.
Le filtre global est appliqué à une séquence de 500 nucléotides comprenant le précurseur au milieu et se base sur le degré de conservation de la structure de la tige-boucle du précurseur prise dans une plus grande séquence.

Les valeurs limites des différents filtres ont été dérivés par apprentissage machine à partir des pré-miARN connus. Pour chaque filtre, la moyenne et la déviation standard sont calculées pour le paramètre considéré. Pour établir la limite, un taux de passage minimal de 90% est choisi, c'est-à-dire que le z-score qui aura permis de retenir 90% des pré-miARN connus sera pris comme limite supérieure. Le choix de cette limite est fait après une analyse visuelle des résultats. Cette limite est devenue nécessaire pour éviter les longs calculs pour peu de résultats car, au-delà de cette limite, il est probable d'obtenir plus de bruit que d'information, ce qui rendrait l'analyse des résultats difficile.

5.3.1 Énergie libre de repliement

L'énergie libre de repliement est le paramètre principale qui permet de mesurer la stabilité des structures secondaires, il est donc impératif d'en tenir compte. Pour ce faire, pour chacun des précurseurs présents dans la base de données, nous avons calculé l'énergie libre de repliement en utilisant le logiciel *RNAfold* de VIENNA-RNA [Hof03]. Si la structure donnée par cet outil ne présente pas une structure en épingle à cheveux, les structures sous-optimales sont considérées pour la recherche d'éventuelles structures valables. Si aucune de ces dernières ne passe le test de la structure, le pré-miARN n'est plus considéré dans l'établissement de ce filtre. Si, par contre, on trouve une structure valide, l'énergie de cette dernière et prise en compte même si elle n'est pas la structure optimale.

L'analyse des résultats montre une distribution de ces derniers principalement entre -35 et -10 kcal/mole. (voir figure 5.3).

Deux structures ont des énergies hors de ces limites : *mir-384* avec une énergie de -6.50 kcal/mol, due à la prédominance de paires de bases A :U et G :U, et *mir-328* avec une énergie de 41.90, kcal/mol due à la prédominance des paires de bases G :C. Des tests ont montré que ces deux valeurs ont une influence minime sur la moyenne et la déviation standard et donc nous les avons gardé dans l'ensemble d'apprentissage.

La moyenne μ des énergies libres des séquences présentes dans notre base de données est calculée par la formule suivante :

$$\mu = \frac{\sum_{i=1}^{N} e_i}{N}$$

où e_i est l'énergie libre de la structure secondaire de la séquence i et N le nombre de séquences dans notre base de données.

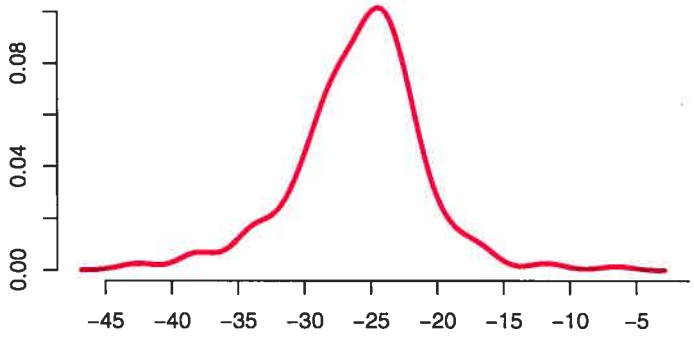

Figure 5.3 – Distribution le l'énergie libre de repliement des pré-miARN. *Les énergies libres de repliement des pré-miARN sont distribuées entre -35 et -10 kcal/mol. La moyenne est de -25.65 kcal/mol et la déviation standard de 4.41.*

Le z-score z_i de l'énergie libre de la structure secondaire de la séquence i est calculé comme suit :

$$z_i = \frac{\mu - e_i}{\sigma}$$

où μ est la moyenne, e_i est l'énergie libre de la structure secondaire de la séquence i et σ est la déviation standard de l'energie libre et est calculée comme suit :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(e_i - \mu)^2}{N}}$$

Pour déterminer le z-score limite, nous avons fait passer tous les précurseurs connus à travers ce filtre d'énergie libre et nous avons relevé le nombre de séquences retenues par le filtre pour certaines valeurs du z-score. Comme le montre la figure 5.4, un taux de réussite de 90% est atteint avec un z-score de 2. Au-delà de cette limite,

les séquences à analyser sont plus nombreuses, impliquant plus de temps de calcul, et sont moins probables d'être des candidates potentielles vu leur éloignement du noyau de la distribution.

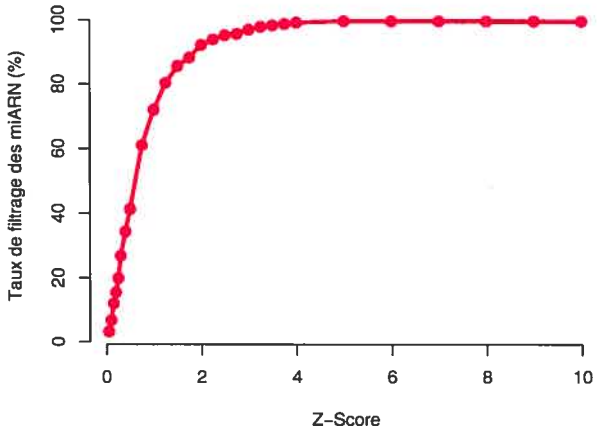

Figure 5.4 – Taux de filtrage des pré-miARN selon le filtre de l'énergie. *Avec un z-score de 2, on peut filtrer 90.93% des pré-miARN connus. Au delà de ce score, l'apport est très marginal.*

5.3.2 Composition nucléotidique des séquences

Le contenu en pourcentage en chacun des nucléotides A, C, G et U est calculé pour les pré-miARN connus. Ce filtre a pour but d'éviter les séquences à forte concentration d'un nucléotide par rapport aux autres car une séquence très riche en GC est susceptible d'avoir une énergie de repliement qui satisfait les exigences même avec une structure ayant très peu d'appariements. Il est vrai que cela n'exclue pas la possibilité qu'elle soit un pré-miARN, mais il ne faut pas perdre de vue que selon les données actuelles ces cas font partie d'un ensemble de cas particuliers sur lesquels il est plus sage de se pencher avec des outils de calcul plus performants.

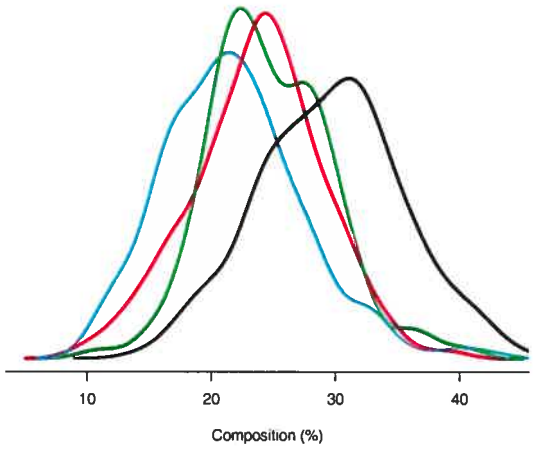

Figure 5.5 – Distribution des compositions nucléiques globales des pré-miARN. *La courbe noire représente la distribution de la composition en U des pré-miARN connus, la courbe rouge représente la composition en A, la courbe verte la composition en G et la courbe cyan la composition en C. Les pré-miARN sont plus riches en U comparativement aux autres nucléotides.*

Ce filtre ne fait qu'une validation sommaire des séquences mais nous trouvons qu'il est indispensable, vue l'impossibilité de faire un alignement multiple de ces dernières. En effet les alignements des séquences des pré-miARN nous donnent des résultats très bruités et difficiles à interpréter, alors à défaut d'utiliser les compositions locales résultant justement de cet alignement multiple, nous avons utilisé les compositions globales. Les distributions des compositions nucléotidiques des pré-miARN montrent que ces derniers sont plus riches en U comparativement aux autres nucléotides. (voir figure 5.5).

La moyenne μ_i de la composition nucléotidique pour le nucléotide i est calculée

comme suit :

$$\mu_i = \frac{\sum_{j=1}^{N} \rho_{ij}}{N}$$

où N est le nombre de séquences dans la base de données et

$$\rho_{ij} = \frac{n_{ij}}{L_j}$$

est le taux de présence du nucléotide i dans la séquence j, n_{ij} le nombre d'occurences du nucléotide i dans la séquence j et L_j la longueur de la séquence j.

Le z-score z_{ij} de la composition nucléotidique pour le nucléotide i de la séquence j est calculée comme suit :

$$z_{ij} = \frac{\mu_i - \rho_{ij}}{\sigma_i}$$

où μ_i est la moyenne de la composition nucléotidique pour le nucléotide i, ρ_{ij} est le taux de présence du nucléotide i dans la séquence j et σ_i est la déviation standard de la composition nucléotidique pour le nucléotide i et est calculée comme suit :

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N} (\rho_{ij} - \mu_i)^2}{N}}$$

Pour déterminer les valeurs des z-score limites pour chacun des nucléotides, le même test a été fait en considérant chacun des filtres. Le tableau 5.1 récapitule les différentes valeurs obtenues pour chacun des 4 nucléotides.

La figure 5.6 montre le taux de passage des pré-miARN connus pour chacun des filtres nucléiques. Nous constatons que l'apport n'est pas conséquent au-delà des z-scores choisis comme limites.

Table 5.1 – z-score limite pour chacun des nucléotides A, C, G et U

Nucléotide	Moyenne	Déviation Standard	Z-score
Adénine (A)	23.79	5.27	1.75
Cytosine (C)	21.81	5.67	1.75
Guanine (G)	24.77	4.90	1.75
Uracile (U)	29.49	4.96	2.25

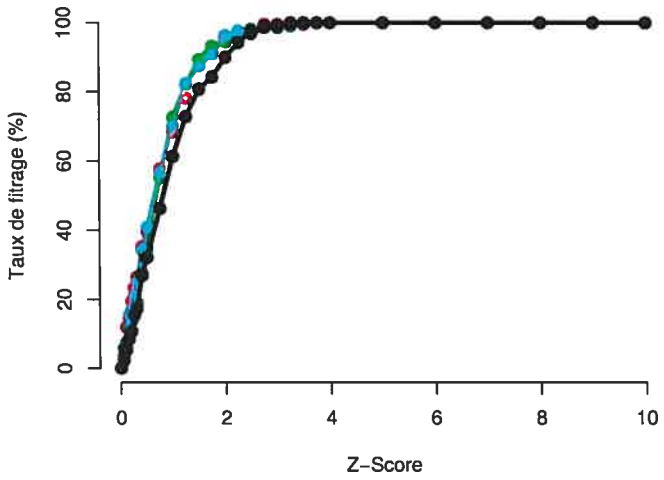

Figure 5.6 – Taux de filtrage des pré-miARN en utilisant les filtres nucléiques

5.3.3 Différence d'énergie libre de repliement

La différence d'énergie libre de repliement est la différence entre l'énergie libre de repliement de la séquence considérée et la moyenne des énergies libres de repliement d'un certain nombre de séquences obtenues par un mélange des nucléotides de celle-ci. Cette différence d'énergie est calculée en utilisant un outil développé au niveau du Laboratoire de Biologie Informatique et Théorique (LBIT) appelé *spf* pour *Structural Pattern Finder*. Cet outil prend en entrée une séquence dont on aimerait identifier les régions présentant un signal de structure. Les principaux paramètres dont nous avons besoins sont la longueur des fenêtres, la longueur du chevauchement entre les fenêtres consécutives et nombre de séquences aléatoires à générer. *Spf* calcule l'énergie libre de repliement de la séquence native et des séquences aléatoires générées par un mélange des nucléotides de la séquence native. La différence d'énergie libre de repliement est la différence entre l'énergie de repliement de la séquence native et la moyenne des énergies libre de repliement des séquences aléatoires. Cet outil est très sensible à la longueur de la fenêtre et au nombre de séquences aléatoires. Pour optimiser le temps de calcul, des tests préliminaires ont été nécessaires pour bien ajuster les différents paramètres. Ainsi, le nombre de séquences aléatoires à partir duquel un changement dans la différence d'énergie de repliement est très peu perceptible est de 100. La longueur de la fenêtre pour laquelle *spf* donne le meilleur pic est de 100 nucléotides comme le montre la figue 5.7. Ces résultats proviennent de l'application de *spf* à des séquences de 500 nucléotides ayant des pré-miARN au milieu. Il est à noter que le temps de calcul pour les 227 pré-miARN de la base de données est de quelques minutes pour les petites fenêtres mais de quelques jours pour des fenêtres de plus de 120 nucléotides.

Une distribution de la différence d'énergie de repliement des pré-miARN a été

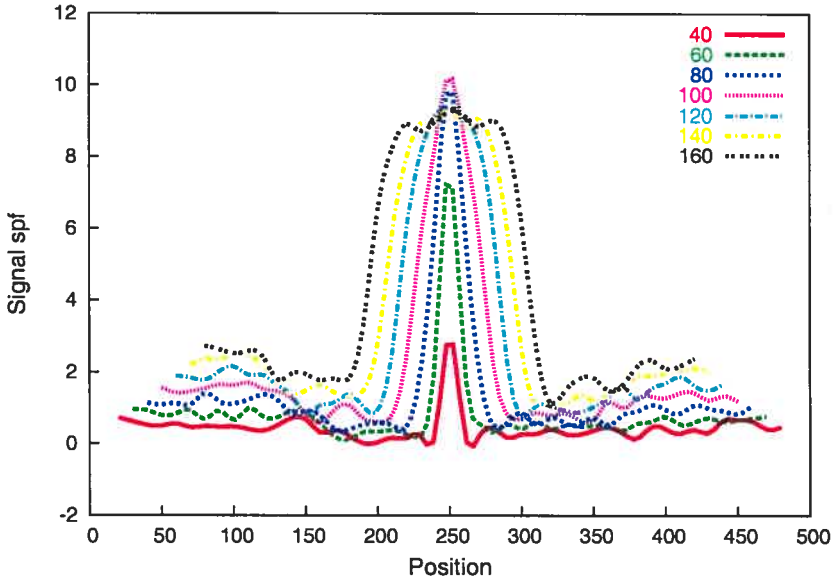

Figure 5.7 – Influence de la longueur de la fenêtre sur le signal de spf. *Le signal spf est plus perceptible pour des séquences d'environ 100 nucléotides.*

établie. Elle montre que les pré-miARN ont des différences d'énergie négatives, comme il a été déjà constaté par Bonnet et ses collègues [BWPR04]. Ceci dénote une certaine stabilité des séquences obtenues étant donnée la composition. Ces différences d'énergie négatives suggèrent que la séquence dont on dispose est parmi les plus stables étant donné la composition nucléotidique. On peut conclure que les différences négatives d'énergies libres de repliement sont une caractéristique des pré-miARN mais qui ne peuvent, malheureusement, être à elles seules un gage de certitude que la séquence représente un bon candidat pour un pré-miARN. A titre d'exemple dans le chromosome 21 de l'humain *spf* a pu trouver plus d'un million de séquences vérifiant ce critère. Comme il est fortement improbable que ce chromosome contienne un aussi grand nombre de miARN, d'autres filtres s'imposent.

Les différents paramètres relatifs à la différence d'énergie libre sont calculés en utilisant les formules servant à calculer les paremètres relatifs à l'énergie libre. Le z-score permettant le filtrage d'au moins 90% des précurseurs est de 1.75 (voir figure 5.8).

5.3.4 Conservation de structure

L'utilisation de la conservation de la structure est motivée par deux principales raisons que nous avons relevé de nos lectures :

1. Les transcrits primaires des miARN sont de quelques centaines de nucléotides de long [LFA93, LKH+04, KW04]. Il est vrai que l'on ne connaît pas avec exactitude ces longueurs car elles sont considérées comme inutiles pour la suite du processus de la biosynthèse mais en prenant des séquences de 500 nucléotides on simule, tant soit peu, ce qui se passe dans la réalité. Cette approche de prendre des séquences de quelques centaines de nucléotides ne limite plus les

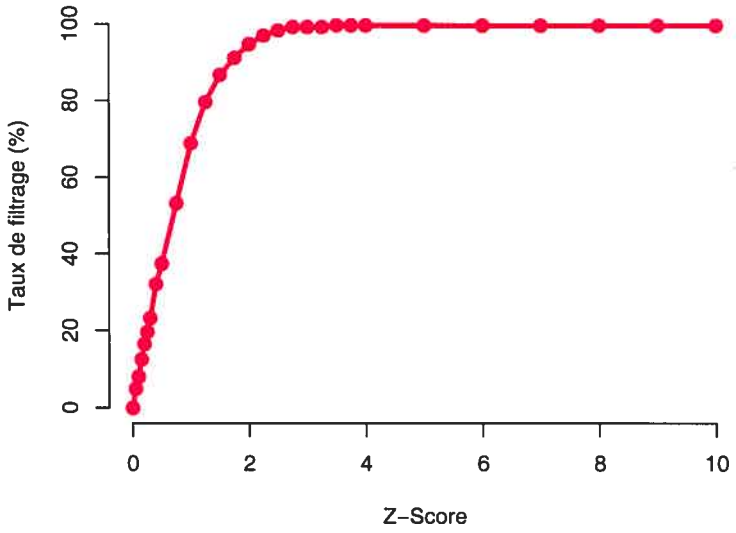

Figure 5.8 – Taux de filtrage des pré-miARN selon le filtre de la différence d'énergie. *Avec un z-score de 1.75 nous arrivons à filtrer 91.47% des pré-miARN connus.*

choix d'appariement des nucléotides à quelques dizaines ou moins. C'est ce qui se passe probablement dans la nature si nous ne considérons pas l'influence des protéines ou autres molécules dans le repliement.

Un fait quasi certain peut être déduit si la structure de tige-boucle n'est pas conservée. En effet, on peut conclure, dans ce cas ci, que la structure de tige-boucle obtenue avec une séquence de quelques nucléotides est le résultat des contraintes indirectes. Avec un peu plus de liberté d'appariement, les nucléotides choisissent d'autres partenaires que ceux qu'on leur a imposés en considérant la petite séquence du précurseur. Il est probable aussi que les nucléotides aux environs de notre précurseur s'apparient de telle sorte qu'une structure de tige-boucle soit impossible à atteindre pour ce dernier, c'est à dire notre précurseur, malgré une composition nucléotidique favorable pour une telle structure.

2. Au lieu de considérer un seul repliement, nous considérons les 100 structures les plus optimales pour les séquences primaires (500 nucléotides) et toutes les structures sous-optimales pour les précurseurs (65 nucléotides). Cela a pour but d'augmenter les chances de trouver la structure naturelle de notre séquence car il est plus probable de trouver la structure naturelle parmi un ensemble de quelques candidats que parmi un ensemble contenant un seul candidat. Le problème que nous risquons de rencontrer concerne la sélection de la structure naturelle. Ce que nous préconisons dans ce cas c'est la considération de la structure la plus conservée dans les structures sous-optimales des séquences de 500 nucléotides.

Pour mesurer le degré de conservation des structures secondaires des pré-miARN, nous avons pris les séquences des précurseurs et des séquences de 500 nucléotides

contenant le précurseur au milieu, que nous appellerons, par abus, séquences primaires. Ces séquences sont repliées en utilisant le logiciel *RNAfold* [Hof03].

Pour le précurseur, toutes les structures sous-optimales sont prises en compte alors que seules les 100 structures ayant les meilleures énergies libres de repliement sont considérées pour les séquences primaires. Nous avons imposé cette limite de 100 structures suite au nombre élevé de structures que nous obtenons, parfois au-delà de 30 000, ce qui nécessiterait beaucoup de temps de calcul lorsque viendra le temps de faire la recherche sur les chromosomes du génome humain.

Après le repliement, chaque structure générée par *RNAfold* [Hof03] pour le précurseur est comparée à la structure que prend cette même séquence dans la structure de la séquence primaire. La vérification se fait de la façon suivante :

Pour chacune des structures du précurseur, nous relevons tous les couples *(i,j)* de nucléotides appariés. Nous faisons de même pour la structure du précurseur dans le repliement de la séquence primaire. Nous localisons le précurseur et nous notons les appariements de chaque nucléotide de celui-ci. Les deux ensembles de couples sont alors comparés et le nombre d'erreurs est calculé. Nous nous sommes permis un jeu d'un nucléotide dans le sens où le i^{eme} nucléotide est permis d'avoir un partenaire à la position *j+1* ou *j-1* dans le but de prendre en compte une des multitudes particularités possibles. Le nombre d'erreurs E_i de la structure i d'un précurseur est calculé selon la formule suivante :

$$E_i = \sum_{j=0}^{N} e_{ij}$$

où N est le nombre de structures générées pour la séquence primaire et e_{ij} le nombre d'erreurs obtenues pour la structure i du précurseur avec la structure j de la séquence primaire. Pour obtenir le taux d'erreurs τ_i dans la structure i on normalise le nombre d'erreurs :

$$\tau_i = \frac{E_i}{bp_i}$$

où bp_i est le nombre de paires de bases dans la structure secondaire i du précurerur considéré. Parmi toutes les structures secondaire sous-optimale du précurseur considéré, on retient celle qui a le plus faible taux d'erreurs. Le taux de conservation ζ_i de la structure secondaire i est : $\zeta_i = 1 - \tau_i$.

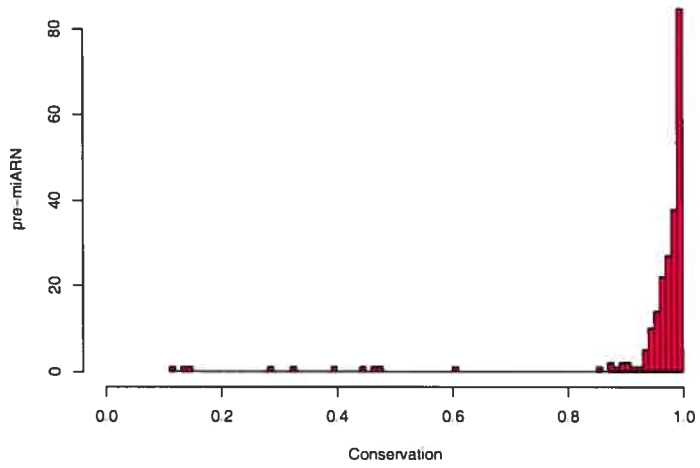

Figure 5.9 – Distribution de la conservation des pré-miARN. *Les pré-miARN ont des structures secondaires très conservées. Au delà de 40% des pré-miARN sont parfaitement conservés.*

Les résultats obtenus montrent que les structures secondaires des pré-miARN peuvent être classées en deux catégories comme illustré dans la figure 5.9. La classe des structures fortement conservées qui comprend plus de 95% du total des séquences et la classe des séquences qui sont peu ou pas du tout conservées. Parmi la classe des structures fortement conservées, on retrouve au-delà de 40% de structures qui

sont parfaitement conservées c'est-à-dire sans aucune erreur.

Pour le calcul de la moyenne de conservation et de la déviation standard utilisées dans ce filtre de conservation, nous avons uniquement utilisé la classe des structures fortement conservées pour les motifs suivants :

- La plupart des miARN peu ou pas conservés sont prédits en se basant sur l'homologie de séquence, ce qui sous-entend qu'il est possible qu'ils ne soient pas de vrais pré-miARN.
- Très peu de pré-miARN ont des structures peu ou pas conservées.
- La considération de tous les résultats nous donnera un filtre permissif. L'utilisation des résultats des structures les moins conservées nous donnera une moyenne faible et une déviation standard plus élevée, ce qui nous rapprocherait un peu du degré de conservation d'autres structures qui ne sont pas forcement des pré-miARN. Ce qui, à notre sens, rendrait ce filtre inutile.

Nous tenons compte aussi de l'asymétrie de la distribution (voir figure 5.9). Puisque les résultats ne sont pas normalement distribués, la moyenne est très peu représentative de l'ensemble. Pour palier à cela, nous avons utilisé le mode de la distribution, qui est 1, au lieu de la moyenne.

Le z-score z_i de la conservation de la structure secondaire de la sequence i est calculé comme suit :

$$z_i = \frac{1 - \zeta_i}{\sigma}$$

où ζ_i est le taux de conservation de la structure de la séquence i et σ est la déviation standard du taux de conservation de la structure secondaire et est calculée comme

suit :

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N}(\rho_{ij} - 1)^2}{N}}$$

où N est le nombre de séquences dans la base de données.

Le z-score limite qui permet filtrer au delà de 90% des pré-miARN est de 2.25.

5.4 Validation

Avant d'appliquer cette nouvelle méthode à quelques chromosomes du génome humain, nous nous sommes donné un petit jeu de tests afin d'avoir une idée de ce que ce filtre est en mesure de faire.

Nous avons crée cinq ensembles de séquences dont les valeurs des différences d'énergies libre de repliement se situent dans la région du z-score permettant de filtrer au minimum 90% des précurseurs connus en considérant ce filtre, soit 1.75. Ces séquences ont des gradients d'énergie de repliement dans la moyenne des différences obtenues avec les pré-miARN connus, ce qui nous fait croire qu'elles présentent des structures avec des degrés de stabilité comparables aux pré-miARN connus. Chacun de ces ensembles contient 250 séquences. Les résultats contenus dans le graphique de la figure 5.10 montrent que peu de ces séquences présentent de la conservation dans leurs structures. En effet pour un z-score de 2.25, nous sommes en mesure de filtrer 90% des précurseurs des microARN connus alors qu'à peine 30% des séquences de tests arrivent à passer le test de la structure. Ceci indique que la conservation de la structure peut être considérée comme une caractéristique des pré-miARN.

5.5 En définitive

Chaque séquence retenue sera pourvue de 7 z-scores, chacun découlant d'une des caractéristiques considérées. Ses z-scores sont convertis en probabilité par la

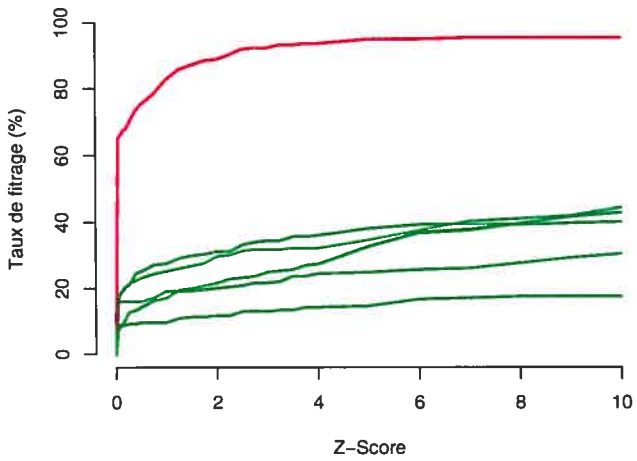

Figure 5.10 – Taux de filtrage par conservation de la structure. *A peine 30% des structures de test passent le filtre de structures pour un z-score de 2.25 alors que 90.35% des pré-miARN passent.*

formule suivante :

$$p_i = \frac{(Z_i - z_i)}{Z_i}$$

où pour une caractéristque i, p_i est la probabilité que la séquence considéré soit un pré-miARN relativement à la caractéristique i, Z_i est le z-score limite considéré et z_i est le z-score obtenu pour la séquence. Pour une séquence donnée, sa probabilité d'être un bon candidat est le produit des probabilités obtenues pour chacune des caractéristiques.

Ces probabilités nous renseignent sur la proximité de chaque séquence analysée du noyau des pré-miARN connus pour une caractéristique donnée. Une probabilité élevée, nous indique une séquence très proche de la moyenne ou du mode donc plus probable d'être un bon candidat.

CHAPITRE 6

RÉSULTATS ET DISCUSSION

6.1 Résultats

Pour valider la méthode proposée, une recherche de candidats de pré-miARN a été réalisée sur des chromosomes humains, pris au hasard, soient les chromosomes 14, 19, 21 et Y. Pour faire des gains en temps de calcul, ces chromosomes ont été divisés en petits fichiers pour partager la recherche sur plusieurs machines. La recherche se fait avec des séquences de 65 nucléotides avec des chevauchements de 55 nucléotides. L'élimination des mauvais candidats se fait au fur et à mesure que les calculs progressent. Seuls les candidats ayant passés tous les filtres sont retenus. Pour le calcul des différences d'énergies libres de repliement données par *spf*, les séquences qui n'auraient pas obtenu une différence négative après 10 mélanges sont rejetées. Le temps de calcul nécessaire pour analyser les quatre chromosomes et d'environ 5 jours. Nous avons utilisé 20 machines dont les processeurs sont de 64 bits et les fréquences d'horloge de 1794 MHz.

De ces quatre chromosomes, on peut avoir plus de 20 millions de candidats possibles en ne considérant que les séquences de 65 nucléotides avec un chevauchement de 55 nucléotides. De cet ensemble de candidats possibles, 91 327 ont été retenus par la méthode soit moins de 0.5%.

Les candidats retenus subissent un prétraitement qui consiste à ne retenir que le meilleur candidat pour chaque portion de 500 nucléotides. D'après ce qui a été constaté dans les pré-miARN connus, la séquence chevauchant en amont ou en aval du pré-miARN possède de bons scores. Par exemple, si le précurseur est situé entre le nucléotide i et le nucléotide j, la séquence située entre $i+10$ et $j+10$ ou

entre *i-10* et *j-10* possède généralement un bon score et est donc retenue par la méthode. Puisque on a pas encore eu de pré-miARN se chevauchant et peu de précurseurs distants de moins de 500 nucléotides, nous avons considéré le candidat le plus probable de chaque ensemble de candidats situés dans une portion de 500 nucléotides. Le prétraitement nous donne 70 002 candidats soit près de 23% des candidats sont disqualifiés.

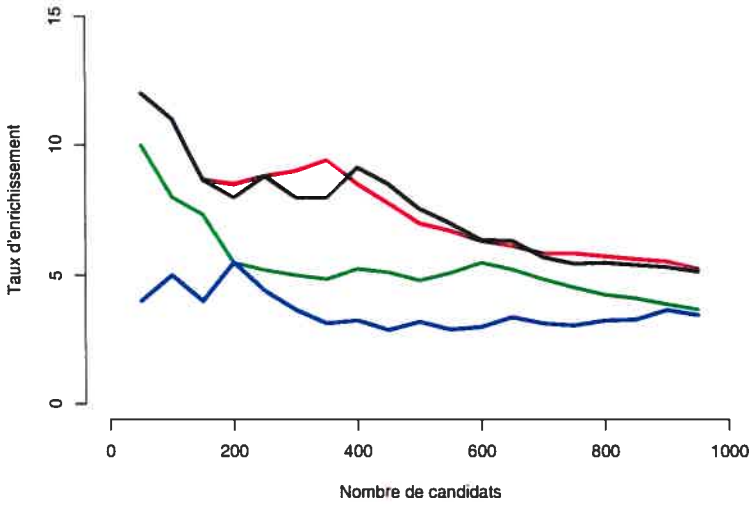

Figure 6.1 – Taux d'enrichissement. *(***Courbe rouge***) Taux d'enrichissement en considérant le chromosome 21 seulement, (***Courbe noire***) chromosomes 21 et Y, (***Courbe verte***) chromosomes 21, Y, et 19 (***Courbe bleue***) chromosomes 21, Y, 19 et 14.*

Pour pouvoir situer les candidats retenus par rapport aux pré-miARN connus, nous avons calculé le taux d'enrichissement pour les quatre chromosomes. Ce taux d'enrichissement est le taux pré-miARN connus dans un certain nombre de candidats retenus par la méthode. Le graphique de la figure 6.1 illustre les résultats pour l'ensemble des quatre chromosomes ajoutés un à un au résultat global.

0.66	AUAUUGCCCAAGCCAGUCUCCAAAUCCUGGGUUCAAGUGAUUUGCUGGCUUGGGACUAUUAGAGU
0.66	UAAGGCUUUCUGCAGAACCAUGCAGCCUCAGAAUCUGCCGAGCUGUGUGUUUCUGUGAAAGUAUU
0.63	CAAUCAAUGAGACCAGUGUCCAGUGUGUAUCAGGAAUGGUGGCAUGCUUUCUAUUACUGGUCUCC
0.53	UAAAAGUGACAUUUGUAGGCAGGCCCAGGCACCUUUGCUUAGGUUUUUGGUCCACAUAUGUCACA
0.53	AAGGGCUGGAGAAAGGAGGUGAUCCCUUCAGCAGCAAAGGGACAUCUCCUUCCUUUUCUUUUAUA
0.53	UAUUUGCUGUAGAGAUUCCCAGUUGUUGUGCUGAAAUGGCUUACACUGGGGAUCCAGCUGCACAA
0.52	AGAGCCUUCUGAAACUUCUCAUGCUGAUGUCUCAGGGGUUUCAGAGGCUCCAUGAUGAAGUUUAU
0.51	UCCUAGACCCUGUUGGCCCAUAAAGUUCCCAGGGUGCCAUAGGGGUCUUGGGAAAUAUUGAAAUU
0.50	UCCUGUAGCAGUGGCCAUUUCUAACCCUAUAAGUAGGGGUAUUAGUUGUAUGGCCCUGCACUGAU
0.50	CUUGUCUCAUACGCUCACAUUUGUGUACAGUACACAAAUGGAGCAUGUGUCUGUGGCAAUCGGCU
0.46	CCUCCCGAGUAGCUGGGAUUACAGGCAAGUGCCACUAUGCCCAGCUAACUUGGUAUUUUUAGUAG
0.45	AGAGACAAGGAAUGGGCUUAGCUGACUUCUCCCAGUAACUUUUGCUGGUUUAUAGCUGAGCCACU
0.45	CUAACUCCUGACGUUGUGAUUUUCGUGCUUUGGACUCCCAAAGUGCUGGAAUUACAAGCAUGAGC
0.44	UUCUUCAUAACCUCUGCUUGCUAGAUAAUUGAGAAGCCUUCUCAGUUGGGGCACAGAGGAGAAUG
0.43	UCUUAUGACUGGGUCCUGCCUACAAGGAGAUUAUACUGUUUCCCUGAUUCAGGACCCAGGUGAUG
0.43	UUCCAUGGGACCUGAUUCUGACACACAUUCUCUUUUGGGAAUGGAGUCAGAACAGCAGUUUCCAG
0.43	CUCAUUACUACCAUAGAGGCACACCUGCUUUCUGACCAAGGUGGUAGGUGGUAGCUUCUGAUUAU
0.43	GUAAUUAUAAUUUGGGGCCUAGGUGUGGUAGCCCACACCUGUAGUCCCAGAACAUUCACUGGCCA
0.43	AUGGUGGAAAGGGUGAGGGAGCUCUCUGGUUCCCUUUUUAUAACGAAACUCAUUCCAUCCAUGAG
0.42	UUUUGAGGGAAAUGGCUGACCCACCUUGUCUUAUCCAUGGUGGAUUAGUAGUCCCUCAAUAGGAA
0.41	AAAGGCUCUCUAACAAACCCAGACUCUUUGGAGUUGGGAGUUGGUUUGCCUGGAACCAACUUC
0.41	UCUAAUGGACAGUGGAGGUUAGUGCAAGAUCUCAGGAUUAUCACUGAGGCCACUGUCCCUCUGUA
0.40	UUCUCCAUCCCAGAAGACUUCUAGAGAGGGUUCCUUUCAAAAUUGGAAGUCCGGGUUCUGGGAGU
0.40	AGCCUUGUUGCUCACACAAAGCCUGUUUGGUGGUCUCUUCACACGGAUGUGUGUGACAAUAAGGU
0.40	UAGCUGCAGUAGCAAAGCCUACACCUGCUCAUUGGUAUGGCUGCUGCUAUGGCAGUAUUAUGGGU
0.39	UUCAACCUAGACAUUGUAGGGGUGAGCCUUUUUGAAACUUACCCCACUGUGAUUUCUAGGUAGAG
0.39	AUGACCAUCACUGUGGCCAUUCCUGCAUUCUGCACAUGCCUGGCUAUAGUGAUAGUGAUAGGAAU
0.38	CAUCUGGAAACUGCCUCUCUCUGGUAUGAGCUGCAAUCAAUUCUCAUUCUAGAGAGGCAGUGUGA
0.37	UUAUGCAGAUGUCAGGAUGUCACUCGGCCCAUUGAAAGUACAAGUGUCACAUACUGGCAUCUGCU
0.37	UGGGAAAGGUAUCUUUAACAACCCCCGACUCUUUGGAGUUGGGAGUGUUGGUUUACCUGGAACCA
0.37	UAACCAGCAUUUGCAUUGCUGGAGCCAUUAGUGAACACUGUGAUGGCCCCAGGAAUGGGCUGAUU
0.36	GAUUUGUAGGGGAGCAGAGGUAGAAUGAUAUGGUUUGCCUCUGUCCCCACCCAAAUCUCAACUUG
0.36	AACCAUGGGGAAACUGCAGGUUAUGGGCUUGAUAGCACAGCCCCUGAUUGCUUUCCAUGGUUAAA
0.36	CCUACCUAGACAUUGUGAGGGUGAGUCUUUUUGAAACUUGCUCCACUAUGAUUUCUAGGUAGAGC
0.36	GUAGCAGUGGCCAUUUCUAACCCUAUAAGUAGGGGUAUUAGUUGUAUGGCCCUGCAAUGAUGGAC
0.36	GAAGGAAUGGUACCAGAGGUACCUUGUACUACCUUGUACCUCUGGUAGAAUUCGGCUGUGAAUCC
0.36	UUCCACCUGUCUUCUCUAUGGUAUCAAGGGGAUUGUCCCAUGACCCUAAGAGAGGGCAGAUGUGU
0.36	CUAUCUAAUUAGCUGGACUACCCACUUCUCAAGGUAAUGGGCCUUGGCAGGCUUAUAGUCCAGUG
0.36	AUACUAUGUUGGGAAGUUACCUGUUAGAGGACUGGGUCCUCUCUUAAGAAGUGACUUCCCAGUAU
0.36	CCUUUUUCAGCAGUGAGGAGAUCUAAACCUUGAUGGUUUUGAAGUGUCACUGCUGCUAAAGAGUC
0.36	UCUCAUCUUGUAGCUCCCAUAAUUCCCAUGUGUUGUGGGAGGGACCUGAUGGGAGAUAACUGAAG
0.36	UGAGAUUUGGGAGGAUCCAGGGUCACAAUAAUAUGGUUUGGUUCUGUGUCCCUACCCAAAUCUUA
0.36	CUGAGGGACAUUGAAGUUCACGUAAUGUCGUCUUCAGUGCCCUCAGGUUCACCUGUACCGUAGAU
0.35	GGUAUGACAGUCAACAUCUCUCCUGUAGGCUGGCUUCAGGGAUGAGAGUACUGUUAUACCUGUGA
0.35	ACAAUGCUCUCUUUAGGCAGGGUCCAGUAAUUAGUGUUACAUUAACUGGGUGCUGGACCCAGCAA
0.35	AGACAUAUCACCUGGCCAAUAGGCUUGGGGGUGUGUGAAAAUAUCCCCUGUUUUCAAGGCCAUUA
0.35	UCCUGGCAAACUCCCUAUUUGAGGGCCUUUAUACCAGGAGCCAAAUGGGAGUGAGGCCAUUUGAU
0.35	AUCAUCUGAAAACUUCCCCAGCUCUGCCUUGAUUGGCUUUAAAUCAGGGAGGGAGAAGGGGACUU
0.34	CAGAGCAGUGCCAUGCCUGUGUCUGAUGCUUUCCAGUAUGGAAAACUGCUCAGAUACUAAUGGUU

Table 6.1 – Liste des 50 séquences présentant les meilleurs scores

La table 6.1 donne les 50 séquences ayant les meilleurs scores données par notre methode. Les pré-miARN connus retenues ne sont pas inclus dans cette liste.

6.2 Discussion

L'analyse plus approfondie des résultats nous a permis de relever les faits suivants. **Beaucoup de faux positifs sont éliminés**. La méthode n'a retenu que 0.5% de toutes les séquences de 65 nucléotides possibles des 4 chromosomes. Malgré que cela n'est pas dû en entier à la conservation de la structure, nous pouvons, néanmoins,

affirmer que la majeure partie des candidats éliminés ne satisfont pas la condition de conservation de structure imposée.

D'autres tests, réalisés sur une partie du chromosome 21 humain, montrent qu'au moins 80% des tiges-boucles sont éliminées pour cause de non conservation de structure. Si l'on tient compte du fait que, pour la même limite 90% des pré-miARN connus sont retenus, nous pouvons conclure que peu de bons candidats soient éliminés par erreur, donc peu de faux négatifs.

En se fiant à la distribution du degré de conservation des pré-miARN connus (voir figure 5.9), on voit bien que la meilleure chose à faire consiste à éliminer les candidats très peu ou pas du tout conservés. En effet, pour les pré-miARN connus près de 65% ont leur structure conservée à plus de 99.95% et que très peu de pré-miARN ont leur structure peu ou pas du tout conservée, ce qui est un bon argument à prendre en compte.

L'autre chose qui pourrait bien influencer les résultats est le contenu de la base de données. Cette dernière contient des séquences qui ne sont pas encore validées par des expériences biochimiques et qui peuvent être de faux pré-miARN comme ce fut déjà le cas auparavant avec *mir-108*. Il est donc possible que les séquences non conservées ne soient pas de vrais pré-miARN. Une petite vérification visuelle de deux pré-miARN (*mir-181c*, *mir-30c-1*) dont le degré de conservation est très faible montre que ces derniers ont été identifiés par homologie de séquence. *Mir-181c* a été prédit, par homologie de séquence à partir du poisson zèbre. Dans cet espèce, la structure du pré-miARN est conservée à près de 83% alors qu'elle est d'à peine 0.05% chez l'humain. Ceci nous mène à penser que se baser sur la séquence du pré-miARN seule et l'homologie de séquence n'est pas assez probant. Nous pouvons donc dire que le manque de spécificité de la méthode est peut-être dû à la présence de faux pré-miARN dans l'ensemble d'apprentissage.

Peu de candidats par portion de 500 nucléotides. Le prétraitement nous permet d'éliminer environ 23% des candidats retenus par la méthode. Ce taux nous indique que nous avons moins de deux candidats dans un segment de 500 nucléotides. Sachant qu'il est possible d'avoir environ 45 candidats possibles sur un segment d'une telle longueur, nous constatons que plus de 90% des séquences de 65 nucléotides possibles ne sont pas retenues. Nous n'avons pas fait de tests pour évaluer les effets de la conservation de la structure par rapport aux autres filtres dans ce cas précis mais tout porte à croire qu'elle a joué un rôle prépondérant.

La même chose a été constatée chez les pré-miARN connus. En concaténant les séquences de 500 nucléotides contenant les pré-miARN connus et en faisant la recherche de candidats par notre méthode nous obtenons 316 candidats de 65 nucléotides avant et 172 après prétraitement dans une séquence de 113500 nucléotides ce qui donne une moyenne de moins de deux candidats par portion de 500 nucléotides.

Très peu de structures moyennement conservées. Sur les 70 002 candidats retenus, 56 356 séquences, soit plus de 80%, ont un très haut degré de conservation. Le même constat peut être fait pour les pré-miARN connus. Très peu de séquences ont des structures moyennement conservées (voir figure 5.9). Cela suggère que la conservation de la structure est peut-être la première caractéristique qu'il faut considérer pour éliminer un grande partie des faux positifs car elle s'applique pour la quasi-totalité des pré-miARN connus. L'assainissement de la base de données, pour ne retenir que les pré-miARN validés par des expériences, donnera une distribution plus représentative de la conservation des pré-miARN. Cela aidera à éliminer le bruit que les faux-vrais pré-miARN aient pu induire.

Des améliorations sont possibles. Le degré élevé de conservation de structure nous indique aussi que ce filtre a atteint ses limites car on ne peut pas aller au-delà de la conservation parfaite. Des améliorations peuvent être apportées à la méthode pour améliorer la conservation des structures qui sont proches de la conservation parfaite car les erreurs peuvent être dues à un décalage dans les appariements plus important que le seul nucléotide considéré dans nos calculs.

En analysant le taux d'enrichissement, on constate que ce dernier est quelque peu faible si l'on ne tient pas compte de la spécificité des scores. En effet, on est tenté de croire que les candidats avec une probabilité élevée sont en quelques sortes trop bons pour être vrais. Mais dans notre cas, ces candidats sont les plus proches du noyau de la distribution ce qui fait d'eux de bons candidats.
C'était prévisible d'avoir un taux d'enrichissement de cette magnitude car les autres critères ne sont pas assez spécifiques aux pré-miARN. En effet, si l'énergie libre de repliement et la différence d'énergie libre de repliement sont relativement discriminantes, la composition nucléotidique, elle, est très générale et n'est utilisée que pour éliminer les séquences avec une prédominance très prononcée d'un nucléotide par rapport aux autres et son apport dans le score final peut être reconsidéré.
Des critères très spécifiques peuvent être intégrés à la méthode pour améliorer sa spécificité. Parmi les critères capables d'apporter un plus, nous avons la composition nucléotidique locale. On a vu que la région porteuse du miARN mature est très riche en U comparativement aux autres régions (voir figure 3.3). Cette caractéristique a été utilisée par Lim et ses collègues dans leur outil de prédiction des pré-miARN chez C. elegans [LLW+03]. Ils ne font état de l'apport de cette caractéristique à la spécificité de leur méthode mais nous considérons que la prédominance de U dans

cette région précise ne peut pas être due au seul fait du hasard.

Dans notre calcul de la conservation, nous nous servons d'une séquence requête de 65 nucléotides que nous replions et nous cherchons la structure résultante dans les structures sous optimales de la séquence de 500 nucléotides. Cette façon de faire ne tient compte que de la structure du précurseur sans se soucier de la conservation de la séquence primaire. Une autre façon de considérer la conservation est de la voir d'une manière plus globale. Il s'agit de prendre la séquence primaire de la replier et de rechercher la sous-structure en tige-boucle la plus stable et la plus conservée parmi toutes les structures sous optimales obtenues pour la séquence primaire.

BIBLIOGRAPHIE

[AH87] V. Ambros and HR Horvitz. The lin-14 locus of caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. *Genes Dev.*, 1(4) :398–414, Jun 1987.

[AJM+05] A. Adai, C. Johnson, S. Mlotshwa, S. Archer-Evans, V. Manocha, V. Vance, and V. Sundaresan. Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res.*, 15(1) :78–91, Jan 2005.

[ALL+05] Y. Altuvia, P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, MJ. Brownstein, T. Tuschl, and H. Margalit. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 33(08) :2697–2706, May 2005.

[AWR91] P. Arasu, B. Wightman, and G. Ruvkun. Temporal regulation of lin-14 by the antagonistic action of two other heterochronic genes, lin-4 and lin-28. *Genes Dev.*, 5(10) :1825–33, Oct 1991.

[Bar04] DP. Bartel. MicroRNAs : genomics, biogenesis, mechanism, and function. *Cell*, 116(2) :281–297, 2004.

[BCG04] MT. Bohnsack, K. Czaplinski, and D. Gorlich. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 10(2) :185–191, Feb 2004.

[BCR05] KM. Brown, CY. Chu, and TM. Rana. Target accessibility dictates the potency of human RISC. *Nat Struct Mol Biol.*, 12(5) :469–70, Jay 2005.

[BHS+03] J. Brennecke, DR. Hipfner, A. Stark, RB. Russell, and SM. Cohen. bantam encodes a developmentally regulated microRNA that

controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113(1) :25–36, Apr 2003.

[BWPR04] E. Bonnet, J. Wuyts, and Y. Van de Peer P. Rouze. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17) :2911–7, Nov 2004.

[CHB04] X. Cai, CH. Hagedorn, and BR. Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10(12) :1957–66, Dec 2004.

[CSD+04] GA. Calin, C. Sevignani, CD. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizuand, S. Rattan, F. Bullrich, M. Negrini, and CM. Croce. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci*, 101(9) :2999–3004, Feb 2004.

[CV01] VL. Chandler and H. Vaucheret. Gene activation and gene silencing. *Plant Physiology*, 125 :145–148, Jan 2001.

[EJ04] S. Eriffiths-Jones. The microRNA registry. *Nucleic Acids Res.*, 32 :Database issue :D109–11, Jan 2004.

[EMP+] SM. Elbashir, J. Martinez, A. Patkaniowska, W. Lendeckel, and T. Tuschl. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate.

[Eps03] RJ. Epstein. Humain Molecular Biology. An Introduction to the Molecular Basis of Health and Disease. 2003.

[FA99] R. Feinbaum and V. Ambros. The timing of lin-4 RNA accumulation controls the timing of postembryonic developmental events in

Caenorhabditis elegans. *Dev Biol.*, 10(1) :87–95, Jun 1999.

[FXM+98] A. Fire, S. Xu, MK. Montgomery, SA. Kostas, SE. Driver, and CC. Mello. Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature*, 391(6669) :806–11, Feb 1998.

[HLY+04] J. Han, Y. Lee, K. Yeom, Y. Kim, H. Jin, and VN. Kim. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev*, 18 :3016–3027, 2004.

[Hof03] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13) :3429–3431, Jul 2003.

[HWR96] I. Ha, B. Wightman, and G. Ruvkun. A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes Dev.*, 10(23) :3041–50, Dec 1996.

[HZ02] G. Hutvagner and PD. Zamore. A microRNA in a multiple-turnover RNAi enzyme complex. *Science.*, 297(5589) :2056–60, Sep 2002.

[Jon02] L. Jones. Revealing micro-RNAs in plants. *Trends Plant Sci.*, 7(11) :473–5, Nov 2002.

[KBT99] R. Kierzek, ME. Burkard, and DH Turner. Thermodynamics of single mismatches in RNA duplexes. *Biochemistry*, 38(43) :14214–23, Oct 1999.

[KSW+04] J. Krol, K. Sobczak, U. Wilczynska, M. Drath, A. Jasinska, D. Kaczynska, and WJ. Krzyzosiak. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem.*, 279(40) :42230–9, Oct 2004.

[KTO+05] Y. Karube, H. Tanaka, H. Osada, S. Tomida ans Y. Tatematsu, K. Yanagisawa Y. Yatabe, J. Takamizawa, S. Miyoshi, T. Mitsudomi, and T. Takahashi. Reduced expression of Dicer associated with poor prognosis in lung cancer patients. *Cancer Sci.*, 96(2) :111–5, Feb 2005.

[KW04] Y. Kurihara and Y. Watanabe. Arabidopsis micro-RNA biogenesis through dicer-like 1 protein functions. *Proc Natl Acad Sci*, 101(34) :12753–12758, Aug 2004.

[KWT96] J. Kim, AE. Walter, and DH. Turner. Thermodynamics of coaxially stacked helixes with GA and CC mismatches. *Biochemistry*, 35(43) :13753–61, Oct 1996.

[LA01] CR. Lee and V. Ambros. An Extensive Class of Small RNAs in Caenorhabditis elegans. *Science*, 294(5543) :862–864, Oct 2001.

[LAH+03] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and VN. Kim. The nuclear RNase III drosha initiates microRNA processing. *Nature*, 425(6956) :415–419, Sep 2003.

[LC04] S. Lu and BR. Cullen. Adenovirus VA1 noncoding RNA can inhibit small interfering RNA and MicroRNA biogenesis. *J Virol.*, 78(23) :12868–76, Dec 2004.

[LDA+05] CH. Lecellier, P. Dunoyer, K. Arar, J. Lehmann-Che, S. Eyquem, and C. Himber ans A. Saib ans O. Voinnet. A cellular microRNA mediates antiviral defense in human cells. *Science*, 5721(308) :557–60, Apr 2005.

[Lew04] B. Lewin. *Genes VIII.* Prentice Hall, 2004.

[LFA93] RC. Lee, RL. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5) :843–854, Dec 1993.

[LGM+05] J. Lu, G. Getz, EA. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, BL. Ebert, RH. Mak, AA. Ferrando, JR. Downing, T. Jacks, HR. Horvitz, and TR. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043) :834–8, Jun 2005.

[LKH+04] Y. Lee, M. Kim, J. Han, KH. Yeom, S. Lee, SH. Baek, and VN. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20) :4051–4060, Oct 2004.

[LKRC02] C. Llave, KD. Kasschau, MA. Rector, and J. Carrington. Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, 14(7) :1605–1619, Jul 2002.

[LLG05] M. Legendre, A. Lambert, and D. Gautheret. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21(7) :841–845, May 2005.

[LLW+03] LP. Lim, NC. Lau, EG. Weinstein, A. Abdelhakim, S. Yekta, MW. Rhoades, CB Burge, and DP Bartel. The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8) :991–1008, Apr 2003.

[LLWB01] NC. Lau, LP. Lim, EG. Weinstein, and DP. Bartel. An abundant class of tiny RNAs with probable regulatory roles in caenorhabditis elegans. *Science*, 294(5543) :858–62, Oct 2001.

[LNP+04] YS. Lee, K. Nakahara, JW. Pham, K. Kim, Z. He, EJ. Sontheimer, and RW. Carthew. Distinct roles for Drosophila Dicer-1 and Dicer-2

in the siRNA/miRNA silencing pathways. *Cell*, 69-81(117) :1, Apr 2004.

[LQRLT01] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of Novel Genes Coding for Small Expressed RNAs. *Science*, 294(5543) :853–858, Oct 2001.

[LXKC02] C. Llave, Z. Xie, KD. Kasschau, and JC. Carrington. Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. *Science*, 297(5589) :2053–2056, Sep 2002.

[MAK02] J. Ma, A.Campbell, and S. Karlin. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol.*, 184(20) :5733–45, Oct 2002.

[MLA97] EG. Moss, RC. Lee, and V. Ambros. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell*, 88(5) :637–46, Mar 1997.

[MLP+04] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell*, 15(2) :185–97, Jul 2004.

[MvdWMM02] MF. Mette, J. van der Winden, M. Matzke, and AJ. Matzke. Short RNAs can identify new candidate transposable element families in Arabidopsis. *Plant Physiol*, 130(1) :6–9, Sep 2002.

[MW05] AA. Millar and PM. Waterhouse. Plant and animal microRNAs : similarities and differences. *Funct Integr Genomics*, 5(3) :129–35, May 2005.

[MYM+05] JB. Ma, YR Yuan, G Meister, Y Pei, T Tuschl, and DJ Patel. Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, 434(7033) :666–70, Mar 2005.

[MYP04] JB. Ma, K. Ye, and DJ. Patel. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature.*, 429(6989) :318–22, May 2004.

[OA99] PH. Olsen and V. Ambros. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol.*, 216(2) :671–80, Dec 1999.

[OAF+05] M. Overhoff, M. Alken, RK. Far, M. Lemaitre, B. Lebleu, G. Sczakiel, and I. Robbins. Local RNA target structure influences siRNA efficacy : a systematic global analysis. *J Mol Biol.*, 348(4) :871–81, May 2005.

[OU05] K. Onishi and S. Ueda. Molecular evolution of a microRNA cluster in the PWS/AS region among mammals. *Gene.*, Mar 2005.

[PAW+03] JF. Palatnik, E. Allen, X. Wu, C. Schommer, R. Schwab, JC. Carrington, and D. Weigel. Control of leaf morphogenesis by microRNAs. *Nature*, 425(6955) :257–63, Sep 2003.

[PL88] WR. Pearson and DJ. Lipman. Improved Tools for Biological Sequence Analysis. *PNAS*, 85 :2444–2448, 1988.

[PRS+00] AE. Pasquinelli, BJ. Reinhart, F. Slack, MQ Martindale, MI Kuroda, B. Maller, DC. Hayward, EE. Ball, Degnan, P.Muller, J. Spring, A. Srinivasan M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the

sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808) :86–9, Nov 2000.

[RGJAB04] A. Rodriguez, S. Griffiths-Jones, JL. Ashurst, and A. Bradley. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, 14(10A) :1902–1910, Oct 2004.

[Rot93] NV. Rothwell. Understanding genetics. a molecular approach. 1993.

[RWR+02] BJ. Reinhart, EG. Weinstein, MW. Rhoades, B. Bartel, and DP. Bartel. MicroRNAs in plants. *Genes Dev.*, 16(13) :1616–26, Jul 2002.

[Sae83] W. Saenger. *Principles of Nucleic Acid Struture.* Springer-Verlag, 1983.

[SJA03] S. Saxena, ZO. Jonsson, and A. Dutta A. Small RNAs with imperfect match to endogenous mRNA repress translation. implications for off-target activity of small inhibitory RNA in mammalian cells. *J. Biol. Chem.*, 278(45) :44312–9, Nov 2003.

[Tam01] W. Tam. Identification and characterization of human BIC, a gene on chromosome 21 that encodes a noncoding RNA. *Gene*, 274 :157–167, Aug 2001.

[Tan05] G. Tang. siRNA and miRNA : an insight into RISCs. *Trends Biochem Sci.*, 30(2) :106–14, Feb 2005.

[TKY+04] J. Takamizawa, H. Konishi, K. Yanagisawa, S. Tomida, H. Osada, H. Endoh, T. Harano, Y. Yatabe, M. Nagino, Y. Nimura, T. Mitsudomi, and T. Takahashi. Reduced expression of the let-7 microR-

NAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.*, 64(11) :3753–6, Jun 2004.

[TS04] A. Tanzer, , and PF. Stadler. Molecular evolution of a microRNA cluster. *J Mol Biol.*, 339(2) :327–35, May 2004.

[VBR+05] A. Vermeulen, L. Behlen, A. Reynolds, A. Wolfson, WS. Marshall, J. Karpilow, and A. Khvorova. The contributions of dsRNA structure to dicer specificity and efficiency. *RNA*, 11(05) :974–682, Apr 2005.

[WC53] JD. Watson and FHC. Crick. Molecular structure of nucleic acids. *Nature*, 171 :737–738, 1953.

[Web05] MJ. Weber. New human and mouse microRNA genes found by homology search. *FEBS*, 272(1), Jan 2005.

[WHR93] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5) :855–62, Dec 1993.

[WRCG31] XJ. Wang, JL. Reyes, NH. Chua, and T. Gaasterland. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol.*, 5(9) :R65, Aug 31.

[XGH04] P. Xu, M. Guo, and B.A Hay. MicroRNAs and the regulation of cell death. *Trends Genet.*, 20(12) :617–24, Dec 2004.

[XJB+98] T. Xia, JJr.SantaLucia, ME. Burkard, R. Kierzek, SJ. Schroeder, X. Jiao, C. Cox, and DH. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42) :14719–35, Oct 1998.

[XVGH03] P. Xu, S.Y. Vernooy, M. Guo, and B.A. Hay. The drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol.*, 13(9) :790–5, Apr 2003.

[ZC04] Y. Zeng and BR. Cullen. Structural requirements for pre-microRNA binding and nuclear export by exportin 5. *Nucleic Acids Res*, 32(16) :4776–85, Sep 2004.

[ZKJ+04a] H. Zhang, F. Kolb, L. Jaskiewicz, E. Westhof, and W. Filipowicz. Single Processing Center Models for Human Dicer and Bacterial RNase III. *Cell*, 118(1) :57–68, Jul 2004.

[ZKJ+04b] H. Zhang, FA. Kolb, L. Jaskiewicz, E. Westhof, and W. Filipowicz. Single processing center models for human Dicer and bacterial RNase III. *Cell*, 118(1) :57–68, Jul 2004.

[Zuk03] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13) :3406–3415, 2003.

[ZYC05] Y. Zeng, R. Yi, and BR. Cullen. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J.*, 24(1) :138–148, Jan 2005.